DEPARTAMENTO DE INTELIGENCIA ARTIFICIAL

Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid

PhD THESIS

Machine Learning in Scientometrics

Author

Alfonso Ibáñez
MS Computer Science
MS Artificial Intelligence

PhD supervisors

Concha Bielza
PhD Computer Science

Pedro Larrañaga
PhD Computer Science

2015

Thesis Committee

President: César Hervás

Member: José Ramón Dorronsoro

Member: Enrique Herrera

Member: Irene Rodríguez

Secretary: Florian Leitner

There are no secrets to success. It is the result of preparation, hard work, and learning from failure.

Acknowledgements

Ph.D. research often appears a solitary undertaking. However, it is impossible to maintain the degree of focus and dedication required for its completion without the help and support of many people. It has been a difficult long journey to finish my Ph.D. research and it is of justice to cite here all of them.

First and foremost, I would like to thank Concha Bielza and Pedro Larrañaga for being my supervisors and mentors. Without your unfailing support, recommendations and patient, this thesis would not have been the same. You have been role models who not only guided my research but also demonstrated your enthusiastic research attitudes. I owe you so much. Whatever research path I do take, I will be prepared because of you.

I would also like to express my thanks to all my friends and colleagues at the Computational Intelligence Group who provided me with not only an excellent working atmosphere and stimulating discussions but also friendships, care and assistance when I needed. My special thank-you goes to Rubén Armañanzas, Roberto Santana, Diego Vidaurre, Hanen Borchani, Pedro L. López-Cruz, Hossein Karshenas, Luis Guerra, Bojan Mihaljevic and Laura Antón-Sánchez. I have learned so much from all of you. Your constructive recommendations and collaboration have been tremendous assets throughout my Ph.D. research.

This dissertation would not have been possible without the financial support offered by the Spanish Ministry of Economy and Competitiveness, projects TIN2008-04528-E and Consolider Ingenio 2010-CSD-2007-00018. I would also like to thank TIN2007-62626, TIN2010-20900-C04-04 and TIN2011-14083-E projects that supported my research during these years.

Finally, my deepest gratitude goes to my wife for her understanding and faithful support as well as her patience and unconditional love. Thanks also for believing in me, being a never-ending fount of moral support and standing by me, when the going is difficult. For all of this, I thank you. Last but of course not least, I wish to thank my parents and sister for their unending encouragement and love from childhood to now. Without them, I would not be where I am. Thank you.

Abstract

Machine learning and scientometrics are the scientific disciplines which are covered in this dissertation. Machine learning deals with the construction and study of algorithms that can learn from data, whereas scientometrics is mainly concerned with the analysis of science from a quantitative perspective. Nowadays, advances in machine learning provide the mathematical and statistical tools for properly working with the vast amount of scientometrics data stored in bibliographic databases. In this context, the use of novel machine learning methods in scientometrics applications is the focus of attention of this dissertation.

This dissertation proposes new machine learning contributions which would shed light on the scientometrics area. These contributions are divided in three parts:

Several supervised cost-(in)sensitive models are learned to predict the scientific success of articles and researchers. Cost-sensitive models are not interested in maximizing classification accuracy, but in minimizing the expected total cost of the error derived from mistakes in the classification process. In this context, publishers of scientific journals could have a tool capable of predicting the citation count of an article in the future before it is published, whereas promotion committees could predict the annual increase of the h-index of researchers within the first few years. These predictive models would pave the way for new assessment systems.

Several probabilistic graphical models are learned to exploit and discover new relationships among the vast number of existing bibliometric indices. In this context, scientific community could measure how some indices influence others in probabilistic terms and perform evidence propagation and abduction inference for answering bibliometric questions. Also, scientific community could uncover which bibliometric indices have a higher predictive power. This is a multi-output regression problem where the role of each variable, predictive or response, is unknown beforehand. The resulting indices could be very useful for prediction purposes, that is, when their index values are known, knowledge of any index value provides no information on the prediction of other bibliometric indices.

A scientometric study of the Spanish computer science research is performed under the publish-or-perish culture. This study is based on a cluster analysis methodology which characterizes the research activity in terms of productivity, visibility, quality, prestige and international collaboration. This study also analyzes the effects of collaboration on productivity and visibility under different circumstances.

Resumen

El aprendizaje automático y la cienciometría son las disciplinas científicas que se tratan en esta tesis. El aprendizaje automático trata sobre la construcción y el estudio de algoritmos que puedan aprender a partir de datos, mientras que la cienciometría se ocupa principalmente del análisis de la ciencia desde una perspectiva cuantitativa. Hoy en día, los avances en el aprendizaje automático proporcionan las herramientas matemáticas y estadísticas para trabajar correctamente con la gran cantidad de datos cienciométricos almacenados en bases de datos bibliográficas. En este contexto, el uso de nuevos métodos de aprendizaje automático en aplicaciones de cienciometría es el foco de atención de esta tesis doctoral.

Esta tesis propone nuevas contribuciones en el aprendizaje automático que podrían arrojar luz sobre el área de la cienciometría. Estas contribuciones están divididas en tres partes:

Varios modelos supervisados (in)sensibles al coste son aprendidos para predecir el éxito científico de los artículos y los investigadores. Los modelos sensibles al coste no están interesados en maximizar la precisión de clasificación, sino en la minimización del coste total esperado derivado de los errores ocasionados. En este contexto, los editores de revistas científicas podrían disponer de una herramienta capaz de predecir el número de citas de un artículo en el fututo antes de ser publicado, mientras que los comités de promoción podrían predecir el incremento anual del índice h de los investigadores en los primeros años. Estos modelos predictivos podrían allanar el camino hacia nuevos sistemas de evaluación.

Varios modelos gráficos probabilísticos son aprendidos para explotar y descubrir nuevas relaciones entre el gran número de índices bibliométricos existentes. En este contexto, la comunidad científica podría medir cómo algunos índices influyen en otros en términos probabilísticos y realizar propagación de la evidencia e inferencia abductiva para responder a preguntas bibliométricas. Además, la comunidad científica podría descubrir qué índices bibliométricos tienen mayor poder predictivo. Este es un problema de regresión multi-respuesta en el que el papel de cada variable, predictiva o respuesta, es desconocido de antemano. Los índices resultantes podrían ser muy útiles para la predicción, es decir, cuando se conocen sus valores, el conocimiento de cualquier valor no proporciona información sobre la predicción de otros índices bibliométricos.

Un estudio bibliométrico sobre la investigación española en informática ha sido realizado bajo la cultura de publicar o morir. Este estudio se basa en una metodología de análisis de clusters que caracteriza la actividad en la investigación en términos de productividad, visibilidad, calidad, prestigio y colaboración internacional. Este estudio también analiza los efectos de la colaboración en la productividad y la visibilidad bajo diferentes circunstancias.

Contents

Part I

Preliminaries

Chapter 1

Introduction

Machine learning is a scientific discipline that addresses how systems can be programmed to automatically learn and to improve with experience. Learning in this context is associated with recognizing complex patterns and make intelligent decisions based on data. The difficulty lies in the fact that the set of all possible decisions given all possible inputs is too complex to describe. To tackle this problem the field of machine learning develops algorithms that discover knowledge from specific data, based on sound statistical and computational principles. In this context, supervised and unsupervised learning methods address issues related to classification, regression, clustering and association problems, among others. Over the past decades machine learning has become one of the mainstays of information technology and with that, an important part of our life.

Data mining can be seen as the application of machine learning to concrete data. It is a step within a larger process, called knowledge discovery in databases. The whole process can be divided into nine steps, including understanding the domain of the problem; generating a dataset; cleaning and preprocessing the data; reducing, projecting and selecting data; identifying the aim of the process; selecting the appropriate methods and algorithms; data mining; interpreting the discovered patterns and, finally, exploiting the new knowledge. Data mining is the main step of the process, and therefore, it is frequently used to refer to the whole process.

In this dissertation, machine learning is used in scientometrics, a field which has grown in popularity during last years. Scientometrics is concerned with the analysis of science from a quantitative perspective. Its major research issues include the measurement of impact, understanding of scientific citations and the production of indicators for use in policy and management contexts. Citation analysis is one of the most widely used scientometric methods. It uses citations in scientific works to establish links to other works or other researchers with the intention of analyzing the frequency, patterns, and graphs of citations in articles. Bibliometric measures have emerged from citation analysis to assess and compare the research activity of individual researchers according to their output. These measures essentially involve counting the number of times scientific papers are cited. They are based on the assumption that influential researchers and important papers will be cited more fre-

quently than others. They constitute an objective method that can summarize the scientific production of a researcher as a set of quantitative figures. Nowadays, many funding agencies and promotion committees use bibliometric measures regularly as a decision-support tool to evaluate almost every research assessment decision. Therefore, the field of scientometrics is hence an increasingly important topic within the scientific community. Finally, scientometrics is not only focused on measuring the literature output but also on analyzing the practices of researchers, the socio-organizational structures, research and development management, the role of science and technology in the national economy, governmental policies towards science and technology, and so on.

As science advances, scientists around the world continue to produce large numbers of research articles. The amount of data that can be not only stored but also processed is getting larger and larger nowadays. Due to the vast amount, human beings cannot directly analyze the information by hand using classical statistical tools. In this context, machine learning provides the tools for properly managing and working with these large amounts of data. Also, it facilitates the mathematical and statistical models that permit to make predictions from experience. It is an important issue because the prediction task could be considered the essence of science.

Chapter outline

This chapter is organized as follows. The main contributions of the dissertation are presented in Section 1.1. Then, the organization of this manuscript is explained in Section 1.2.

1.1 Contributions of the dissertation

Based on the motivation that machine learning could provide the tools for properly working with the vast amount of scientometrics data, this dissertation presents different contributions which would shed light in the prominent area of scientometrics and pave the way for new applications. These contributions are presented in three parts.

Predicting bibliometric indices: One of the most commonly employed measures of professional recognition is the number of times an article is cited by fellow researchers. Although using citations to judge the quality of journals papers has been criticized, it should be noted that citations frequently correlate with other forms of professional recognition like winning a Nobel Prize. Consequently, citations will serve as a proxy for the professional recognition received by a journal article. Nowadays, publishers of scientific journals face the tough task of selecting high quality articles that will attract as many readers and citations as possible from a pool of articles. In this context, the first part of the dissertation proposes several predictive models to forecast the citation count of an article within the first four years after publication. The possibility of a journal having a tool capable of predicting the citation count of an article before it is published would pave the way for new assessment systems.

Beyond traditional measures like the number of citations, one of the most successful bibliometric measures is the h-index which combines both the quantity and visibility of researcher's publications into a single-number criterion. This indicator has received a lot of attention from researchers over the last few years since it is used by funding agencies and promotion committees to evaluate the importance of research. Considering the popularity of the h-index, several cost-sensitive models are also proposed to predict the annual increase for a four-year time horizon. These new models are not interested in maximizing classification accuracy, but in minimizing the expected total cost error derived from mistakes in the classification process. The use of models capable of predicting the h-index that a researcher will have in coming years could be a useful tool for the scientific community.

Discovering new associations among indices: Many bibliometric indices have been developed in the literature to take into account aspects not previously covered. The result is that, nowadays, the diversity of bibliometric indices poses the challenge of exploiting the relationships among them. In this context, the second part of this dissertation deals with analyzing relationships among bibliometric indices by means of Bayesian network models. The proposed models analyze the joint probability distribution over all analyzed indices and discover new conditional (in)dependencies relationships among triplets of indices. Also, they perform all kinds of probabilistic reasoning, and measure how some indices influence others in probabilistic terms.

Besides discovering new relationships among indices, researchers have also turned their attention to the predictive power of bibliometric indices in many situations. Therefore, scientific community now faces the challenge of selecting which of this pool of bibliometric indices have a higher predictive power. In this context, a method for identifying a core set of bibliometric indices for prediction purposes, i.e., relevant indices which have a higher predictive power, is also proposed. This method solves a proposed multi-output regression problem where the role of each variable is unknown beforehand. Gaussian Bayesian networks and genetic algorithms are used to select which subset of bibliometric indices best corresponds to predictive variables and which group can be considered as response variables. The resulting predictive indices are very useful for prediction purposes, that is, when the relevant index values are known, knowledge of any index value provides no information on the prediction of other bibliometric indices.

Exploring Spanish computer science research: National exercises for the evaluation of scientific research are becoming regular events in ever more countries. In general, these exercises are aimed at informing selective funding allocations, stimulating better research performance, and demonstrating that investment in research is effective and delivers public benefits, among others. Until recently, the conduct of these evaluation exercises has been founded on the so-called peer-review methodology, where research products submitted by scientists are evaluated by appointed panels of experts. In general, these assessments give the greatest weight to output quality. But recent developments in scientometrics, particularly

for measurement of publication quality, have lead many policy-makers to introduce the more or less extensive use of bibliometric indicators in their research assessments. In this context, the third part of this dissertation carries out a scientometric analysis of the computer science field in Spain using bibliometric indices.

The proposed scientometric analysis is achieved at macro (nationwide), meso (universities) and micro (researchers) levels. It provides a comprehensive overview of the current situation of scientific production under the publish-or-perish culture. It is commonplace that the pressure to publish has affected researchers' behavior in the sense that it is not only important what they write, but also how often, where and with whom they write. Therefore, an overview is required to characterize research activity and analyze how the publish-or-perish culture affects Spanish computer science research. Also, the third part of the dissertation presents a robust cluster analysis methodology to analyze universities and their academic staff and identify both their strengths and weaknesses in terms of productivity, visibility, quality, prestige and international collaboration. Finally, the effects of collaboration on productivity and visibility is also studied under different circumstances.

1.2 Overview of the dissertation

The manuscript includes 13 chapters grouped into five parts:

Part I. Introduction

This is the current part.

- Chapter 1 introduces this dissertation, stating the main contributions of the dissertation and summarizing the document organization.

Part II. Background

This part includes three chapters introducing the basic concepts and definitions used throughout this dissertation. The chapters explain the basic theory behind the models and tools used in the following chapters. The state-of-the-art is discussed in each of these chapters.

- Chapter 2 presents an overview of machine learning. The main machine learning approaches, supervised learning and unsupervised learning, are discussed in depth. The different methods used in this dissertation are briefly reviewed, and some notes are given on how to evaluate the performance of the machine learning methods.

- Chapter 3 introduces probabilistic graphical models, with a special focus on Bayesian networks. The chapter includes the theoretical foundations of Bayesian networks in discrete and continuous domains, and discusses some of the issues that will be addressed during this dissertation, e.g., parameterization, learning from data and inference. Also, specific Bayesian network models for solving supervised learning problems are reviewed.

- Chapter 4 includes an introduction to scientometrics. Citation analysis and bibliometric indices are presented as scientometric methods to quantify science, technology and innovation. The well-known h-index and other measures are reviewed to assess scientific research. Also, a comparison of the main features of most important bibliographic databases (Web of Science, Scopus and Google Scholar) is presented.

Part III. Data Mining in Research Evaluation

This part consists of three chapters including data mining proposals in research evaluation. The objective of this part is two-fold. First, several supervised predictive models are learned from data to forecast bibliometric indices values like the number of citations and the h-index. Second, the relationships among bibliometric indices are analyzed by Bayesian models which discover probabilistic conditional (in)dependencies among triplets of indices. Finally, the chapters explain all steps from data acquisition to evaluation of obtained results.

- Chapter 5 presents a tool capable of predicting the citation count of a journal article before it is published. In this context, several predictive models (naive Bayes, logistic regression, and decision trees, among others) are learned for the Bioinformatics journal. To build these models, tokens found in the abstracts of Bioinformatics papers have been used as predictive features, along with other features.

- Chapter 6 incorporates cost-sensitive learning and feature subset selection into new predictive models. These models are used to forecast the annual increase of the h-index for Neurosciences journals in a four-year time horizon using a set of bibliometric indices. The proposed models are not interested in maximizing classification accuracy, but in minimizing the expected total cost error derived from the classification process.

- Chapter 7 analyzes how bibliometric indices relate (irrelevant, dependent and so on) to each other by means of Bayesian network models. The induced Bayesian networks are then used to discover probabilistic conditional (in)dependencies among the indices and, also for probabilistic reasoning. A case study of 14 well-known bibliometric indices on computer science and artificial intelligence journals is performed to test the reliability of proposed models. Using these models, editorial boards could answer many questions related to their journal citation indices.

Part IV. Exploring Spanish Computer Science Research

This part includes five chapters which analyze the Spanish computer science research. The first three chapters achieve a comprehensive overview of the current situation of the Spanish computer science research under the publish-or-perish culture. In contrast, the last two chapters focus on building different models to predict the scientific success of Spanish computer science academics and to uncover the best core set of relevant indices which have a higher predictive power.

- Chapter 8 presents an overview of the Spanish computer science research, including different analysis at the macro, meso and micro levels. Parameters such as number of documents, number of citations, number of citations per document, number of authors per document, document types, types of collaboration, and computer science disciplines, are analyzed in this chapter. Finally, a comprehensive overview of the current situation in the area of computer sciences is achieved.

- Chapter 9 develops a cluster analysis methodology for measuring the performance of research activities in terms of productivity, visibility, quality, prestige and internationalization. It permits a robust cluster analysis whose results can be used to characterize the Spanish computer science research activity of universities and academic staff, identifying both their strengths and weaknesses. Also, this methodology could support policy-makers in the processes of strategic planning, in verifying the effectiveness of policies and initiatives for continuous improvement.

- Chapter 10 analyzes the relationship among research collaboration, number of documents and number of citations, that is, how publication and citations vary by number of authors. These measures are also analyzed under different circumstances, i.e., when documents are written in different types of collaboration, when documents are published in different document types, when documents are published in different computer science subdisciplines, and, finally, when documents are published by journals with different impact factor quartiles.

- Chapter 11 deals with the prediction of scientific success of Spanish computer science academics. An approach based on cost-sensitive Bayesian classifiers forecasts the annual increase of the h-index for a four-year time horizon using some author-based variables (area, position, university, seniority) and 12 bibliometric indices. The proposed model takes into account the expected cost of instances predictions at classification time.

- Chapter 12 deals with the challenge of exploiting the relationships among bibliometric indices. This chapter uncovers the best core set of relevant indices which have a higher predictive power for forecasting other bibliometric indices. This results in a novel multi-output regression problem where the role of each variable (predictor or response) is unknown beforehand. Gaussian Bayesian networks and genetic algorithms are used to solve the above problem and discover new multivariate relationships among indices.

Part V. Conclusions

This part concludes this dissertation.

- Chapter 13 summarizes the contributions of this dissertation and the scientific results derived from it. The chapter also discusses the research lines opened in this work and summarizes future research topics.

Part II

Background

Chapter 2

Machine Learning

2.1 Introduction

The advances in computer technology has enabled the possibility of storing vast amount of data. Despite this, it is not feasible to analyze this information using classical statistics tools. Therefore, machine learning provides the tools for managing and working with these data.

Most data acquisition devices are now digital and record gigabytes of data every day. For example, a supermarket chain has many stores selling thousands of products to millions of customers. The point of sale terminals are able to record the details of each transaction. These stored data becomes useful only when it is analyzed and turned into information that the supermarket can make use, for example, to discover relationships among products and to make predictions about sales and stocks. In this retail context, the consumer behavior is not completely random. People do not go to supermarkets and buy things at random. When they buy beer, they buy chips; they buy ice cream in summer and hot drinks in winter, and so forth. Such patterns can be found using machine learning techniques and may help supermarkets understand the consumer behavior.

Machine learning is inherently a multidisciplinary field which draws on results from research fields as diverse as: artificial intelligence, Bayesian methods, computational complexity theory, control theory, information theory, philosophy, psychology, and neurobiology [327]. It is concerned with building systems which automatically learn programs from data and make accurate predictions without human intervention. It also rests upon the theoretical foundation of statistical learning theory which provides conditions and guarantees for good generalization of learning algorithms [458].

Some authoritative definitions of machine learning follows: Arthur Samuel, one of the pioneers in the field, defined machine learning as a "field of study that gives computers the ability to learn without being explicitly programmed". Tom Mitchell [327] also stated that "the field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience". Finally, Ethem Alpaydin [19] provided a definition for machine learning as "programming computers to optimize a performance criterion using example data or past experience".

Machine learning can be used to solve different problems:

- Classification: It focus on identifying to which of a set of labels a new observation belongs, on the basis of a set of data containing observations whose label membership is known. Although the set of labels could have several discrete values, binary classification is probably the most frequently studied problem in machine learning. This can be thought of as a discrimination problem, modelling the differences or similarities between labels. An example would be the spam detection. Given an email inbox, the goal is to identify those email messages that are spam and those that are not. Having a model of this problem would allow a program to leave non-spam emails in the inbox and move spam emails to a spam folder.

- Regression: It is concerned with modelling the relationship between variables labelled with real values rather than labels. More specifically, it helps one understand how the value of the output variable changes when any one of the predictor variables is varied, while the other predictor variables are held fixed. An example would be predicting the price of a used car based on car attributes like brand, year, engine capacity and mileage, among others.

- Clustering: It is based on grouping a set of observations in such a way that observations in the same group are more similar to each other than to those in other groups. Usually there is more than one way of partitioning the data into meaningful groups, therefore there is no right or wrong answer. In this case, observations are not labelled with labels, but can be divided into groups based on similarity and other measures of natural structure in the data. An example would be face detection, that is, organizing pictures by faces without names. Given a digital photo album of many hundreds of digital pictures, the aim is to identify those photos that include a given person. A model of this decision process would allow a program to organize photos by person.

- Associations: It is intended to discovering interesting relations between variables in large databases. An example would be the product recommendation. Given a purchase history for a customer and a large inventory of products, the objective is to identify those products in which that customer will be interested and likely to purchase. A model of this decision process would allow a program to make recommendations to a customer and motivate product purchases.

The above problems, and many others, are solved using different machine learning algorithms. These algorithms are usually categorized into supervised and unsupervised methods. Supervised methods infer a mapping function from a set of labeled data. They rely on a set of observations for which the target property is known. These methods are trained on this set of observations, and the resulting mapping is applied to further observations for which the target property is not available. In contrast to supervised methods, unsupervised methods try to find hidden structures in unlabeled data. This is an advantage in the sense that the data can

speak for themselves without preconceptions such as expected classes being imposed. Since the observations given to these methods are unlabeled, there is no error signal to evaluate a potential solution. Finally, the categorization of machine learning algorithms into these two groups is not a concluding division. Other categories, like semi-supervised learning methods [84, 490], reinforcement learning methods [434] and deep learning methods [40], also cover machine learning algorithms but they are out of the scope of this thesis.

Chapter outline

On the one hand, Section 2.2 introduces an overview of supervised learning approaches. It also shows the measures and methods used to estimate how well classifiers predict the class value for new instances. On the other hand, Section 2.3 provides an overview of unsupervised learning approaches and their main dissimilarity measures. Finally, it also focus the clustering validation problem, including internal and external validity indices.

2.2 Supervised learning

Supervised learning [45, 130] is the most widely studied approach in machine learning. It addresses the problem of predicting the class of a new observation based on a set of features describing its main properties. For instance, an application of supervised learning is the credit concession. Financial institutions usually calculate the risk that a customer could pay the loan back given some predictive features, e.g., the amount of credit and the information about the customer (income, savings, profession, age, past financial history, and so forth). In this context, a supervised learning approach could be able to calculate the risk for a new application and then decides to accept or refuse it accordingly.

The objective of supervised learning is to build models based on training data, and then, predict testing data using the learnt model. The training data must be characterized using pairs of descriptive features and a class label variable, whereas the test data are characterized using only the descriptive features. Let the training set $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, c^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, c^{(N)})\}$ be a set of instances described by a tuple of a vector of features, that is, $\boldsymbol{x}^{(i)}=\{x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}\}$, and a label from a class variable $c^{(i)} \in \Omega(C)=\{c_1, c_2, \ldots, c_k\}$, then a supervised classification algorithm builds a model, learnt from $\mathcal{D}$, which will be used to assign class labels to new instances, $\{\boldsymbol{x}^{(N+1)}, \boldsymbol{x}^{(N+2)}, \ldots, \boldsymbol{x}^{(N+M)}\}$.

2.2.1 Supervised learning approaches

Many supervised learning algorithms have been developed in the literature [264]. They can be organized into different classification approaches such as Bayesian classifiers, decision trees, instance-based learning, regressions, kernel methods, neural networks and ensemble learning, among others. Although the scope of this thesis does not cover all mentioned approaches, a brief introduction to the most important approaches is carried out.

Bayesian network classifiers Bayesian network classifiers [43, 163] are a class of Bayesian networks specially designed to solve supervised classification problems. These classifiers model the joint probability distribution over the predictive variables and the class variable. The Bayes rule [37] is used to classify a new instance $\boldsymbol{x}$ according to its predictive variables. The class with the maximum posterior probability is selected as the class label for the available instance. Bayesian network classifiers will be studied in depth in Section 3.5.

Decision trees Decision tree learning [61, 339, 341] is one of the most used and practical approach for inductive inference. Decision trees are hierarchical models that represents the knowledge of the problem with a tree structure by a recursive division of the predictive variables' space. Their goal is to build tree structures whose nodes are as pure as possible, that is, they contain observations of a single class value. Each node in the decision tree specifies a test of some variable of the problem, and each branch descending from that node corresponds to one of the possible values for this variable. The leaf nodes of a decision tree contain the class label assigned to each subregion of the problem. Once the decision tree structure has been created, a new instance is classified by starting at the root node of the tree structure. Then, the new instance moves down the tree branch corresponding to the value of the variable specified by this node. This process is repeated for the sub-tree rooted at the new node as long as it takes to reach the appropriate leaf node, then returning the class label associated with this leaf.

Several algorithms can be used to construct a tree based on some data set. For example, the ID3 and C4.5 algorithms [372] are greedy search algorithms that construct a tree recursively and choose at each step the variable to be tested using the information gain ratio, so that the separation of the data examples is optimal. The C4.5 algorithm is an extension of ID3 and made a number of improvements to ID3: C4.5 deals with both continuous and discrete variables, it handles variables with missing values and different costs, and it could go back through the tree structure and attempts to remove unnecessary branches by replacing them with leaf nodes. Although algorithms belonging to this approach are interpretable, efficient and reasonably accurate, they have problems like overfitting, among others [372].

Instance-based learning Instance-based learning approaches [14, 112, 299] do not provide an explicit model as the other paradigms when training examples are provided. Instead, they simply store the training examples until a new instance to be classified appears. Given a new instance, its relations to the already stored examples are examined in order to assign a class label value for the new instance. This approach classify a new instance by looking for the most similar instances in the training dataset and returning their labels. If the most similar instances have different class labels, combination rules have been proposed [13].

Examples of instance-based learning include k-nearest neighbor classifiers [103, 196] and locally weighted regression methods [25]. On the one hand, the simplest method is arguably the k-nearest neighbor classifier. Here, the k points of the training data closest to the test point are found according to some distance metric, and a class label is then assigned depending

on the class labels for the k closest training instances. This class label is usually given to the test point by a majority vote between the k points. On the other hand, locally weighted regression performs a regression around a point of interest using only training data that are local to that point. Methods belonging to this approach are highly intuitive and attains, given their simplicity, remarkably low classification errors. In contrast, these methods are computationally expensive and require a large memory to store the training data.

Linear models Linear models [346, 375] are excellent and simple methods for classification and numeric prediction. They have been widely used in statistical applications for decades. These models are easy to understand: the final output is usually a weighted sum of the input variables. The magnitude of the weight shows the importance of each variable and its sign indicates if the effect is positive or negative. Of course, these methods suffer from the disadvantage of linearity since if the data exhibits a nonlinear dependency, the resulting solution may not fit data very well.

Linear regression [406, 474] is the most appropriate method when the class is numeric, and all the variables are also numeric. This technique expresses the class as a linear combination of the attributes with predetermining weights as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \ , \tag{2.1}$$

where y is the class, $x_1, x_2, \ldots, x_n$ are the attribute values, and $\beta_0, \beta_1, \ldots, \beta_n$ are weights calculated from the training data. The sum of the squares of the difference between the predicted and the actual values over all the training instances is minimized to compute the coefficients β. This technique can easily be used for classification in domains with numeric attributes. The trick is to perform a regression for each class value, setting the output equal to one for training instances that belong to the class value and zero for those that do not. The result is a linear expression which approximates a numeric membership function for each class value. Then, given a test example of unknown class, calculate the value of each linear expression and choose the one that is largest.

Although linear regression often yields good results in classification, it has some drawbacks [478]. First, the membership values it produces are not proper probabilities because they can fall outside the range 0 to 1. Second, least squares regression assumes that the errors are not only statistically independent, but are also normally distributed with the same standard deviation, an assumption that is blatantly violated when the method is applied to classification problems because the observations only ever take on the values 0 and 1.

Logistic regression [214, 322] does not suffer from the above problems. It is used to predict the class of new instances in a binary classification problem by using a linear function of the predictive features as

$$p(C = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}} \tag{2.2}$$

$$p(C=0|\boldsymbol{x}) = 1 - p(C=1|\boldsymbol{x}) = \frac{e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}} \tag{2.3}$$

where $\beta_0, \beta_1, \ldots, \beta_n$ are the parameters of the model. The estimation of these parameters is based on the maximum likelihood estimation method. These parameters describe the size of the contribution of each variable to the model.

Kernel methods Kernel-based algorithms [337, 394, 413] provide a simple bridge from non-linearity to linearity problems. These methods use a linear classifier to solve a non-linear problem by mapping the original non-linear observations into a higher-dimensional space where the problem is easier to model. It is hoped that the data could become more easily separated in this new higher-dimensional space. This approach was first used to solve binary classification problems by means of support vector machines [58, 106, 459].

Support vector machines work by mapping the training data into a feature space by the aid of a kernel function that computes a similarity between two given observations. Using this transformation, the problem becomes linearly separable and can be solved using decision hyperplanes [4, 69]. Although their training time is very high, and, the learnt model cannot be easily interpreted, they have yielded very accurate results and are less prone to overfitting than other methods. Also, the attractiveness of such classifiers stems from their elegant treatment of nonlinear problems and their efficiency in high-dimensional problems.

Neural networks Artificial neural networks [44, 201, 315, 378] are computational models that tries to simulate the structure and/or functional aspects of biological neural networks. These networks consist of an interconnected group of processing units, also called neurons, and processes information using a connectionist approach to computation. They are composed by one or more layers of processing units connected with each other. Each processing unit aggregates the inputs that it could receive from the environment or could be the outputs of other processing units, and sends the result, which is a weighted sum of the inputs in the simplest case, to other processing units. The connection between processing units are modeled with weights.

The simplest networks are called perceptrons [382, 383] which have a single layer of processing units. Although these classifiers are able to distinguish labels in a binary classification problem by means of a linear discrimination function, they present some limitations [326]. If a set of instances is not linearly separable, perceptrons will never reach a model where all instances are classified properly. Multi layered perceptrons have been created to try to solve this problem [387]. These networks are usually used to model complex relationships between inputs and outputs by means of nonlinear discrimination functions. There are several algorithms with which a network can be trained [345]. However, the most well-known and widely used learning algorithm to estimate the values of the weights is the back propagation algorithm [387], which use gradient descent to tune network parameters to best fit a training set of input-output pairs. Finally, artificial neural networks usually provide higher accuracies than other methods. However they operate as a black box and they are difficult to interpret.

Cost-sensitive algorithms Cost-sensitive algorithms [140, 486] are not interested in maximizing classification accuracy, but in minimizing the expected total cost error derived from mistakes in the classification process. They take into account matrices of misclassification cost to express relative distances between classes. This approach incorporates decision-making costs to define fixed and unequal misclassification costs between classes. The cost model takes the form of a cost matrix, where the cost of classifying a sample from a true class j in class i corresponds to the matrix entry m_{ij}. The diagonal elements of this matrix are set to zero, meaning correct classification has no cost.

Cost-sensitive algorithms can be divided into two main categories. Algorithms belonging to the first category (direct methods) [128, 294, 442] design classifiers that are naturally cost-sensitive, using directly the misclassification costs in the learning algorithms. Most of the works belonging to this category are devoted to make decision trees cost-sensitive. A detailed survey of cost-sensitive decision trees induction algorithms can be found in [297]. In contrast, the second category (indirect methods) convert existing cost-insensitive classifiers into cost-sensitive classifiers. These classifiers can be further categorized into relabeling methods, weighting methods and sampling methods. Specially, relabeling methods [126, 478] relabel the classes of training or testing instances by applying the minimum expected cost criterion [266]. This criterion is defined by fixed misclassification costs and posterior probabilities as follows: $R(c|\boldsymbol{x}) = \sum_{c' \in \Omega(C)} p(c'|\boldsymbol{x})\ cost(c|c')$. Weighting methods [436] assign a weight to each instance in terms of its class according to misclassification costs, that is, instances, which carries a higher misclassification cost, are assigned proportionally high weights. Finally, sampling methods [414, 487] modify the class distribution of training data according to their costs and then directly apply cost-insensitive classifiers on the sampled data.

2.2.2 Classification validation

The evaluation of the performance of classifiers is a matter of on-going debate among researchers [422]. It is a key step in any supervised learning problem since its aim is to estimate how well a classifier predicts the class value for new instances. The validation of supervised classifiers is relatively simple procedure due to the presence of real class values. Based on these real class values, a classifier correctly classifies a new instance if the predicted class is the same as than the real class, that is counted as a success; if not, it is an error.

The confusion matrix [432] is an important tool to validate the performance of classifiers. It is a specific table layout that allows visualization of the performance of a classifier. In a dichotomic classification problem, each column of the matrix represents how many instances have been classified as been either positive or negative, while each row represents how many of those classifications were according to the real class value and how many were not. The main diagonal values in the confusion matrix correspond to the corrected classified instances, which are the number of true positive (TP) and the number of true negatives (TN). The missclassification values are divided into false negatives (FN) and false positives (FP).

Real / Predicted	as Positive	as Negative
Positive	TP	FN
Negative	FP	TN

Performance measures Supervised learning has several ways of evaluating the performance of classifiers that they produce. Several measures of the quality of classification can be directly obtained using these values from the confusion matrix. For classification problems, it is natural to measure a classifier's performance in terms of the error rate. The classifier predicts the class of each instance: if it is correct, that is counted as a success; if not, it is an error. The error rate is the proportion of errors made over a whole set of instances, and it measures the overall performance of the classifier. It is also usually expressed in terms of the classification accuracy, that is, the proportion of success made over a whole set of instances. The accuracy of a classifier is thus the probability of correctly classifying a new instance, and is estimated by

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.4}$$

The total number of correctly classified instances is usually the score to be maximized in the most general ways of comparing algorithms. When a dataset is unbalanced, the above score is not representative of the true performance of that classifier because it does not distinguish between the number of correct labels of different classes. In this context, two measures that separately estimate a classifier's performance on different classes are:

$$sensitivity = \frac{TP}{TP + FN} \tag{2.5}$$

$$specificity = \frac{TN}{FP + TN} \tag{2.6}$$

The sensitivity estimates the probability of the positive label being true, that is, the ratio of positive instances that are correctly classified as positive. The counterpart of sensitivity for the negative instances is the specificity. It estimates the probability of the negative label being true, that is, the ratio of negative instances that are correctly classified as negative. Both measures distinguish the correct classification of labels within different classes. Finally, other measures have been developed to address situations in which the cost of a false positive and the cost of a false negative are very different. An example of these measures is precision which can be defined as the probability that an instance classified as positive is actually positive. From the confusion matrix, it is computed as

$$precision = \frac{TP}{TP + FP} \tag{2.7}$$

Estimation methods Besides defining performance measures, it is necessary to defined methods to honestly estimate these measures. There are several approaches in the literature. Resubstitution [417] is a simple estimation method which consists of training a classifier with the full dataset and testing its performance again on the same whole dataset. It is not likely to be a good method since presents a high bias due to the specific data and, thus, provides accuracy estimations which are too optimistic. Although it cannot be considered an honest method for estimating any performance measures, resubstitution can be useful as an upper bound of accuracy performance. With the intention of solving the previous optimism problem, hold-out [278] splits the dataset into two disjoint sets, one to induce the model and the other one to estimate its performance. In this way, an honest estimation is achieved by training the classifier with a set of instances that are independent from the ones used to testing. The disadvantage of this method is that a subset of instances is not used to build the model, which cannot be desirable if the available data sample size is not very high.

The most frequently used evaluation method is called k-fold cross validation [433]. This method solves the possible disadvantage of hold-out by means of dividing all instances from the dataset into k randomly disjoint subsets of approximately equal size. Each subset is used to test a model that is learned from the other k-1 subsets. The k accuracy values are averaged to output the estimated value of the model learned from all instances. This procedure can be repeated several times to reduce the variance of the estimate, giving rise to the repeated k-fold cross validation [258]. Then, the final estimate of the accuracy is the mean of the estimates computed in each repetition. Another improvement of the original k-fold cross validation is the stratified k-fold cross validation [61] which tries to preserve the proportion of instance of each class in every fold. This method, which obtains more realistic estimations, is recommended when the class labels are imbalanced.

Another popular evaluation technique is the leave-one-out method [333]. It is a special case of k-fold cross validation where the number of folds is equals to the number of instance in the training set, that is, the learning process is repeated k times, using k-1 instances to learn, and using a single instance to test each time.

The above presented performance measures and methods are accepted and frequently used by the machine learning community. There are more performance measures and methods in the state of the art of classification validation. Of remarkable relevance could be the area under the ROC curve [427] and the F-measure [455] as performance measures, and jackknife [376] and bootstrap [132] as methods to estimate the performance measures.

2.3 Unsupervised learning

Unsupervised learning is the most frequently analyzed machine learning problem after supervised learning. It studies the problem of finding groups of similar observations in a dataset. In the case of a company with customer data (demographic information and past transactions with the company), the company may want to see the distribution of the profile of its customers. In this context, a clustering model allocates customers similar in their attributes to

the same group, providing the company with natural groupings of its customers. Once such groups are found, the company may decide specific strategies to different groups. Outliers can be also detected by the clustering model, so they may imply a niche in the market that can be further exploited by the company.

Clustering is concerned with finding a structure in a collection of unlabeled elements that are characterized by several variables. The goal is to group elements in this collection so that elements that belong to a cluster are very similar to each other, whereas different clusters are highly heterogeneous.

A formal definition of clustering is as follows. Let the data set $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ be a set of instances described by a vector of descriptive features in a space of dimension F, that is, $\boldsymbol{x}^{(i)} \in \Re^F, \forall i \in \{1, \ldots, N\}$. In this way, the goal is to assign a cluster label $c^{(i)}$ to each instance, with $c^{(i)} \in \{1, \ldots, K\}$, based on some similarity measure with the other instances. The final number of clusters, K, is often unknown and must be estimated.

Dissimilarity measures Clustering approaches usually rely on the definition of a distance or dissimilarity measure between the observations. These measures play an important role in clustering approaches, like partitional or hierarchical, since their results can be completely different according to the selected dissimilarity measure. Many dissimilarity measures can be found [125]. One of the most used dissimilarity measure is the Euclidean distance. For instance, the distance between two instances is calculated with the Euclidean distance as

$$Euclidean(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(i+1)}) = \sqrt{\sum_{j=1}^{F} \left(x_j^{(i)} - x_j^{(i+1)}\right)^2} \quad . \tag{2.8}$$

The Euclidean distance is a concrete case of the general Minkowski distance:

$$Minkowski(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(i+1)}) = \sqrt{\sum_{j=1}^{F} w_j^r \left(x_j^{(i)} - x_j^{(i+1)}\right)^r} \quad , \tag{2.9}$$

where w_j is a possible weight for feature j, and r the distance norm. Manhattan distance is another measure following this structure, with r=1. There are other different measures, like the Mahalanobis distance [306] or Pearson's correlation [360], based on correlations between features. The definition of the Mahalanobis distance is as follows

$$Mahalanobis(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(i+1)}) = \sqrt{(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i+1)})^T \Sigma^{-1} (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i+1)})} \quad , \tag{2.10}$$

where Σ is the covariance matrix of features in a space of dimension F.

2.3.1 Unsupervised learning approaches

Different starting points and criteria usually lead to different taxonomies of clustering algorithms [143, 234, 481]. A simple agreed frame is to classify clustering techniques as partitional clustering, hierarchical clustering and probabilistic clustering, based on the properties of the

clusters generated. Partitional clustering groups elements exclusively, so that any element belonging to one specific cluster cannot be a member of another cluster. On the other hand, hierarchical clustering produces a hierarchical structure of clusters. Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones (agglomerative clustering) or by splitting larger clusters (divisive clustering). Finally, probabilistic clustering provides a cluster membership probability for each element, where elements have a specific probability of being members of several clusters. The above clustering techniques are somehow used throughout this thesis. An introduction of these techniques are detailed as follows.

Partitional clustering Partitional clustering algorithms assign a set of instances into K pre-fixed number of clusters with no hierarchical structure. In principle, the optimal partition, based on some dissimilarity measure, can be found by enumerating all possibilities. But this brute force method is infeasible in practice, due to the expensive computation [295]. Therefore, heuristic algorithms have been developed in order to seek approximate solutions.

The k-means algorithm [303] is a well-known partitional clustering algorithm. This algorithm process as follows. Fisrt, the K clusters are initialized [355], obtained K centroids, which represent the K cluster centers, being μ_k the centroid of cluster C_k. Each instance $\boldsymbol{x}^{(i)}$ is assigned to a cluster by minimizing the distance between the instance and cluster centroids. One each instance is assigned to a cluster, the cluster centroids are recalculated based on those assignments. After the new centroids are calculated, the instances are again reallocated in the clusters. This is an iterative process that converges when cluster centroids do not suffer any changes from an iteration to another. This algorithm works conveniently only with numerical attributes and can be negatively affected by a single outlier. Some outliers, which are quite far away from the cluster centroid, are still forced into a cluster and, thus, distorts the cluster shapes. In this way, new algorithms, like Partitioning Around Medoids (PAM) [249], have appeared in order to overcome these obstacles.

PAM begins by selecting an instance as a medoid for each cluster C_k. After selecting a set of K medoids, K clusters are constructed by assigning each instance to its nearest medoid. If the objective function can be reduced by switching a selected medoid for an unselected (non-medoid) element, then they are switched. This continues until the objective function can be decreased no further. This algorithm has several advantages with regard to K-means. First, this algorithm presents no limitations on attributes types because it utilizes real data points (medoids) as the cluster prototypes (medoids do not need any computation and always exist). Second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore, it is lesser sensitive to the presence of outliers. Unlike k-means, the resulting clustering is independent of the initial choice of medoids. The objective of this algorithm is to determine a representative element (medoid) among the elements of the dataset for each cluster. For K clusters, the goal is to find K representative instances which minimize the following objective function

$$E = \sum_{k=1}^{K} \sum_{\boldsymbol{x}^{(i)} \in C_k} d(\boldsymbol{x}^{(i)}, m_k), \tag{2.11}$$

where K is the number of clusters, $\boldsymbol{x}^{(i)}$ is an instance belonging to the cluster C_k, m_k is the medoid of cluster C_k, and $d(\boldsymbol{x}^{(i)}, m_k)$ is the dissimilarity measure between $\boldsymbol{x}^{(i)}$ and m_k.

Hierarchical clustering Hierarchical clustering also rely on the definition of a dissimilarity measure between the instances. Hierarchical clustering algorithms build a tree of clusters called dendrogram. This dendrogram allows exploring data on different levels of granularity. Hierarchical clustering algorithms are categorized into agglomerative and divisive [233]. Agglomerative clustering starts with clusters and each of them includes exactly one instance. A series of merge operations based on the linkage function are then followed out that finally lead all instances to the same group. Divisive clustering proceeds in an opposite way. It starts with one cluster of all instances and recursively splits the most appropriate objects. For a cluster with N instances, there are $2^{N-1} - 1$ possible two-subset divisions, which is very expensive in computation [143]. Therefore, divisive clustering is not commonly used in practice. Despite this, some divisive clustering algorithms, like MONA and DIANA [249], are also developed in the literature.

Based on the different definitions for distance between two clusters, there are many agglomerative clustering algorithms. The simplest methods include single linkage [420] which calculates the distance between the two closest instances in each cluster, and complete linkage technique [426] which calculates the distance between the two remote instances in each cluster. Other linkage metrics, such as median linkage, centroid linkage and average linkage [421], are also developed.

Unlike methods based on linkage metrics, a more complicated clustering algorithm called the Ward's method [471] uses an analysis of variance approach to evaluate the distances between clusters. It is also known as Ward's minimum variance method. Given K clusters, the Ward's algorithm reduces them to $K - 1$ mutually exclusive clusters by considering the union of all possible $K(K-1)/2$ pairs. It selects the union of clusters which minimizes the heterogeneity among cluster elements. Thus, homogeneous clusters are linked to each other. The complete hierarchical structure can be obtained by repeating this process until only one cluster remains.

Finally, hierarchical clustering techniques do not generate a single partition but a hierarchy of clusters. Different dissimilarities measures and linkages functions yield different hierarchies of clusters. Therefore, these decisions should be carefully made taking into account the nature of the data. Also, a limitation of hierarchical clustering is that divisions in the divisive, and mergers in the agglomerative paradigm, cannot be undone once made.

Probabilistic clustering Probabilistic clustering deals with the problem of fitting a finite mixture of distribution [317], where each component is the probability distribution that models the observations belonging to the cluster. Although other distributions can be used,

the most popular mixture model is formed by Gaussian components [318]. In this way, each cluster k is represented by one component $f_k(\boldsymbol{x})$ of the mixture. Each distribution (k) is characterized by two parameters for each variable (j): the mean (μ_{kj}) and the standard deviation (σ_{kj}).

Using this approach, the clustering problem becomes a mixture parameter estimation problem. Once the parameters are estimated, they can be used to calculate the posterior probabilities of each instance and distribution. The parameter estimation is performed using methods such as AutoClass [85] or SNOB algorithms [467], but the most widely used is the Expectation-Maximization (EM) algorithm [124, 316].

The EM algorithm is an iterative method that is used to find the maximum likelihood estimates of the mixing coefficients (π_k) and the parameters of the conditional Gaussian distributions (μ_{kj} and σ_{kj}). Thus,

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{k=1}^{K} \pi_k \; f_k(\boldsymbol{x}) \\ &= \sum_{k=1}^{K} \pi_k \; \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \\ &= \sum_{k=1}^{K} \pi_k \prod_{j=1}^{F} \left[\frac{1}{\sqrt{2\pi}\sigma_{kj}} e^{-\frac{1}{2}\left(\frac{x_j - \mu_{kj}}{\sigma_{kj}}\right)^2} \right]. \end{aligned} \tag{2.12}$$

The algorithm converges to a locally optimal solution by iteratively updating values for π_k, μ_{kj} and σ_{kj}. This whole process is embedded in a cross-validation procedure that is capable of estimating the number of clusters K without this having to be set a priori.

2.3.2 Clustering validation

One of the most important issues in cluster analysis is the evaluation of clustering results [194]. Clustering validation is concerned with checking the quality of clustering results and determining the optimal number of clusters (the best for the input dataset) by means of quality measures [324]. Recent works [22, 190] have focused on the comparison of cluster validity indices. It is usual to classify these indices into two groups (internal and external validity indices) but the classification criteria are not always clear [233, 484].

Internal validation Internal validity indices do not require a priori information from the dataset, they are based on the information intrinsic to the dataset alone. The optimal number of clusters is usually determined based on internal validity indices. These indices are used to measure the goodness of a clustering structure (compactness and separation of the clusters) without external information. The internal validity indices used throughout this thesis are well-known in the literature. Some of the most widely used indices are detailed as follows.

Silhouette index [385] measures cluster cohesion using the distance between all the points in the same cluster and the cluster separation using the nearest neighbour distance. A larger

average silhouette coefficient indicates a better overall quality of the clustering result. For an instance $x^{(i)}$, it is defined as

$$Silhouette\ (\boldsymbol{x}^{(i)}) = \frac{b(\boldsymbol{x}^{(i)}) - a(\boldsymbol{x}^{(i)})}{max(b(\boldsymbol{x}^{(i)}), a(\boldsymbol{x}^{(i)}))} \quad , \tag{2.13}$$

where $a(\boldsymbol{x}^{(i)})$ is the average dissimilarity between instance $\boldsymbol{x}^{(i)}$ and all other points in the cluster where $\boldsymbol{x}^{(i)}$ belongs, and $b(\boldsymbol{x}^{(i)})$ is the minimum average dissimilarity to instance of each different clusters.

Davies-Bouldin index [113] estimates the cluster cohesion based on the distance from the points in a cluster to its centroid and the cluster separation based on the distance between centroids. A lower Davies-Bouldin index indicates better clustering. It is calculated by averaging each pair of clusters as

$$Davies\text{-}Bouldin = \frac{1}{K} \sum_{k=1, k \neq k'}^{K} max \left(\frac{d_k + d_{k'}}{d(\mu_k, \mu_{k'})} \right) \quad , \tag{2.14}$$

where K is the total number of clusters, d_k and $d_{k'}$ are the average distances of all instances in each cluster to their respective centroid μ_k and $\mu_{k'}$. Finally, $d(\mu_k, \mu_{k'})$ is the distance between cluster centroids.

Calinski-Harabasz index [75] estimates the cluster cohesion based on the distances from the points in a cluster to its centroid, whereas the cluster separation is based on the distance from the centroids to the global centroid. A high index value indicates isolated and unified clusters. This index can be defined as

$$Calinski\text{-}Harabasz = \frac{BSS_K(K-1)}{WSS_K(N-K)} \quad , \tag{2.15}$$

where BSS_K is the between-cluster sum of squares, WSS_K is the within-cluster sum of squares, K is the total number of clusters and N is the total number of instances.

External validation External validity indices require previous knowledge about dataset to check the quality of clustering results. When the correct partition of a dataset is available the usual approach is to compare it with the partition proposed by the clustering algorithm. It is based on one of the many indices that compare the agreement between two different data partitions. Given a set of N instances and two different partitions, S and T, to be compared, then, a is defined as the number of pairs of instances that are located in the same group in S and in T, b is the number of pairs of instances in the same group in S but not in T, c is the number of pairs of instances in the same group in T but not in S, and d is the number of pairs of instances in different groups in both partitions S and T. Given this context, some of the most widely used external validity indices are presented as follows:

On the one hand, the Rand index [374] is defined as

$$Rand = \frac{a+d}{a+b+c+d} \quad . \tag{2.16}$$

This index lies between 0 and 1. It takes the value of 1 when the two clusterings are identical. The problem of this index is its value when two random partitions are compared, since it does not take a zero value.

On the other hand, the adjusted Rand index [216] is defined as

$$Adjusted\ Rand = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{N}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]} \quad . \tag{2.17}$$

This index overcomes the Rand index limitation concerning the random partitions. It introduces a penalization to avoid the possibility of random classification. This index also lies between 0 and 1, the latter being the value output when two partitions are equals.

Finally, the Russel index [388] is defined as

$$Russel = \frac{a}{a+b+c+d} \quad . \tag{2.18}$$

This index only considers pair of instances in the same group in both partitions as good decisions. This index also lies between 0 and 1. A high index value indicates two partitions are highly similar.

Chapter 3

Probabilistic Graphical Models

3.1 Introduction

Probabilistic graphical models [81, 259, 280, 477] have received many attention from machine learning community over the last years, due to their capability for knowledge discovery and reasoning under uncertainty. They combine probability theory and graph theory into a single framework that is able to manage many real-world problems. These models are composed of two main components: a graphical component and a probabilistic component. The graphical component is a graph where the nodes represent the variables in the problem domain and the edges represent the conditional (in)dependence relationships among the variables, whereas the probabilistic component models these dependence relationships using (conditional) probability distributions. The main characteristic of these models is that they consist of a graphical structure and a set of parameters that together encode a joint probability distribution for the variables in the problem domain.

The basket analysis in a supermarket chain can be an application of probabilistic graphical models. This application tries to learn associations between products bought by customers. The main idea is to learn a conditional probability of the form $p(Y|X)$ where Y is the product conditioned on X, which is the set of products which the customer has already purchased.

Although different probabilistic graphical approaches have been introduced in the literature such as Bayesian networks [237, 356, 357], Markov networks [256] and chain graphs [282], among others, this thesis is focused on Bayesian networks, as it is the most frequently used model for reasoning with uncertainty in many problems [366].

Chapter outline

Section 3.2 introduces Bayesian network models and their parameterizations according to the nature of variables. Section 3.3 provides an overview of the state-of-the-art approaches dealing with Bayesian network structure and parameter learning. Section 3.4 describes probabilistic inference in Bayesian networks, which consists of estimating the posterior probability of some variables of interest given evidence of the value of some other variables in the Bayesian

network. Finally, Section 3.5 deals with Bayesian network classifiers, a class of Bayesian networks for solving supervised learning problems.

3.2 Bayesian networks

A Bayesian network [203, 238, 280] is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph. Formally, a Bayesian network is defined as a pair $(\mathcal{S}, \boldsymbol{\theta})$. The first element, $\mathcal{S}$, is a directed acyclic graph, $\mathcal{S} = (\mathcal{V}(\mathcal{S}), \mathcal{A}(\mathcal{S}))$, with a set of nodes given by the random variables of the problem, i.e., $\mathcal{V}(\mathcal{S}) = \{X_1, \ldots, X_n\}$, and a set of arcs $\mathcal{A}(\mathcal{S}) \subseteq \mathcal{V}(\mathcal{S}) \times \mathcal{V}(\mathcal{S})$ representing the probabilistic conditional (in)dependencies among the nodes. The second element, $\boldsymbol{\theta}$, is a vector of conditional probabilities that in combination with $\mathcal{S}$ allows the factorization of the joint probability distribution over $(X_1, \ldots, X_n)$ as:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid \boldsymbol{\Pi}(X_i)), \tag{3.1}$$

where $\boldsymbol{\Pi}(X_i)$ represents the set of parents of X_i. A node X_i is a parent of another node X_j if there is an arc from X_i to X_j. The probabilistic component, $\boldsymbol{\theta}$, determines the kind of probability distributions used in a Bayesian network. Different parameterizations of Bayesian networks have been proposed depending on the nature of the random variables.

3.2.1 Discrete Bayesian networks

In the discrete domain, the statistical relationship between a variable X_i and its parents $\boldsymbol{\Pi}(X_i)$ is encoded using discrete probability distributions which are defined by conditional probability tables. These tables store the parameters of the discrete probability distributions of each variable for all the combinations of the values of its parents, that is, $\theta_{ijk} \equiv p(X_i{=}x_i^{(j)} \mid \boldsymbol{\Pi}(X_i){=}\pi(x_i)^{(k)})$ where $x_i^{(j)}$ is the jth value of variable X_i and $\pi(x_i)^{(k)}$ is the kth

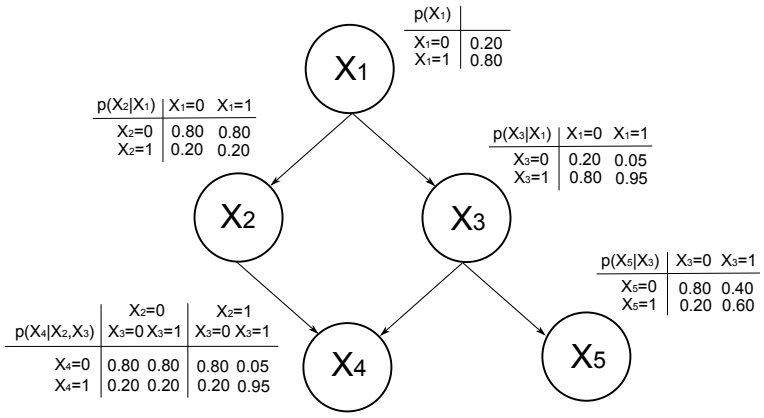


Figure 3.1: Bayesian network example: graphical and probabilistic components

combination of values of the parents of X_i. Figure 3.1 illustrates an example of a Bayesian network where all the variables are binary. It is observed that variable X_1 has no parents, whereas $\mathbf{\Pi}(X_2)=\{X_1\}$, $\mathbf{\Pi}(X_3)=\{X_1\}$, $\mathbf{\Pi}(X_4)=\{X_2, X_3\}$ and $\mathbf{\Pi}(X_5)=\{X_3\}$. Thus, the Bayesian network in Figure 3.1 encodes the following factorization of the joint probability distribution:

$$p(\boldsymbol{X}) = p(X_1)\ p(X_2|X_1)\ p(X_3|X_1)\ p(X_4|X_2, X_3)\ p(X_5|X_3) \tag{3.2}$$

3.2.2 Gaussian Bayesian networks

A Bayesian network is said to be a Gaussian Bayesian network [178, 411] if and only if its associated joint probability distribution is a multivariate normal distribution, $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, with a joint probability density function

$$f(\boldsymbol{x}) = (2\pi)^{-n/2}|\mathbf{\Sigma}|^{-1/2} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{3.3}$$

where $\boldsymbol{x}$ is a realization of the random variables, $\boldsymbol{\mu}$ is the n-dimensional mean vector, $\mathbf{\Sigma}$ is the $n \times n$ covariance matrix, $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$, and $\boldsymbol{\mu}^T$ is the transpose of $\boldsymbol{\mu}$.

The joint probability distribution of the variables in a Gaussian Bayesian network can be specified as in Equation (3.1) by the product of a set of conditional probability distributions

$$f(X_i \mid \mathbf{\Pi}(X_i)) \sim \mathcal{N}\left(\mu_i + \sum_{x_j \in \Pi(X_i)} \beta_{ij}(x_j - \mu_j), v_i\right), \tag{3.4}$$

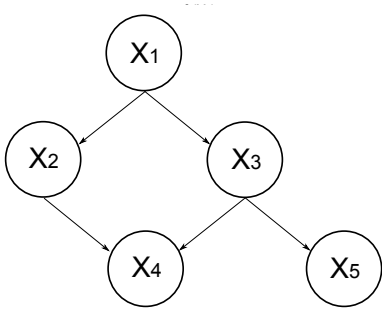
where μ_i is the unconditional mean of X_i, β_{ij} is the regression coefficient of X_j in the regression of X_i on its parents $\mathbf{\Pi}(X_i)$, and v_i is the conditional variance of X_i given its parents. It can be calculated as

$$v_i = \Sigma_{X_i} - \mathbf{\Sigma}_{X_i\mathbf{\Pi}(X_i)}\mathbf{\Sigma}^{-1}_{\mathbf{\Pi}(X_i)}\mathbf{\Sigma}^T_{X_i\mathbf{\Pi}(X_i)}, \tag{3.5}$$

where Σ_{X_i} is the unconditional variance of X_i, $\mathbf{\Sigma}_{X_i\mathbf{\Pi}(X_i)}$ is the row matrix with covariances between X_i and $\mathbf{\Pi}(X_i)$, and $\mathbf{\Sigma}_{\mathbf{\Pi}(X_i)}$ is the covariance matrix of $\mathbf{\Pi}(X_i)$. Finally, Figure 3.2 shows an example of Gaussian Bayesian network structure and its joint probability distribution.

3.3 Learning Bayesian networks

The learning Bayesian network problem [203, 344] can be divided into two tasks: structural learning, that is, to identify the topology of the Bayesian network, and parametric learning, that is, to estimate the conditional probabilities for a given Bayesian network topology. Both the structure, $\mathcal{S}$, and the parameters of the probability distributions, $\boldsymbol{\theta}$, can be obtained in two ways. The first way uses expert knowledge for the learning task, whereas the second way uses algorithms which learn Bayesian networks from a dataset $\mathcal{D}$. Learning Bayesian



$$f(\boldsymbol{X}) = f(X_1)f(X_2|X_1)f(X_3|X_1)f(X_4|X_2,X_3)f(X_5|X_3)$$

where
$f(X_1) \sim \mathcal{N}(\mu_{X_1}, v_{X_1})$
$f(X_2|X_1) \sim \mathcal{N}(\mu_{X_2}+\beta_{X_1X_2}(x_1\text{-}\mu_{X_1}), v_{X_2})$
$f(X_3|X_1) \sim \mathcal{N}(\mu_{X_3}+\beta_{X_1X_3}(x_1\text{-}\mu_{X_1}), v_{X_3})$
$f(X_4|X_2,X_3) \sim \mathcal{N}(\mu_{X_4}+\beta_{X_2X_4}(x_2\text{-}\mu_{X_2})+\beta_{X_3X_4}(x_3\text{-}\mu_{X_3}), v_{X_4})$
$f(X_5|X_3) \sim \mathcal{N}(\mu_{X_5}+\beta_{X_3X_5}(x_3\text{-}\mu_{X_3}), v_{X_5})$

Figure 3.2: Gaussian Bayesian network structure and its joint probability distribution

networks from expert knowledge [150, 176] is out of the scope of this dissertation. Therefore, the remaining of the section presents a brief review on principled approaches for learning Bayesian networks from data. This is a very active field of research, and there have been several new proposals in the last years [68, 215, 460, 483].

3.3.1 Structural learning

The most difficult task in Bayesian networks is to determine their structure $\mathcal{S}$, that is, which node should be connected to which node. The task of automatically defining the structure from a dataset is called Bayesian network structure learning. There are two basic approaches for learning the structure of a Bayesian network: algorithms based on constrained methods [115, 358, 431] and score+search methods [98, 204]. Constraint-based methods use conditional independence tests to identify the dependent and independent relationships among variables and then build a directed acyclic graph. In contrast, score+search methods approach the structure learning problem as an optimization problem. They use a search procedure to explore the space of network structures, and a score function to evaluate the candidate network structures and guide the search procedure.

Constraint-based methods Constraint-based methods [87, 418] perform statistical tests to determine a large percentage of the conditional (in)dependence relationships among the variables in the given dataset and then a directed acyclic graph is built. The PC algorithm [430] is a well-known example of constraint-based methods. It starts from a complete undirected graph, then performs recursive conditional independence tests for deleting edges. The result is a skeleton in which all edges are still undirected and should be transformed into arcs using edge orientation rules. Some improved version of the PC algorithm [67, 246] have been also developed in the literature. A major weakness of methods belonging to this constraint-based approach is that too many tests may have to be performed, with each test being built upon the results of another. This may lead to escalated errors in structure identification. Additionally, increasing cardinality in the conditioning part dramatically reduces test reliability.

Score+search methods Most of developed structure learning algorithms fall into the score+search category. This approach states the learning task as an optimization problem, and two main components (a scoring metric and a search strategy) have to be defined. Once a score metric is specified, a search method is needed to move in an intelligent way through the space of possible directed acyclic graphs and find the structure with the optimal score. The fitness score measures the quality of every candidate structure with respect to a dataset. The number of candidate structures that can be built from data grows more than exponentially as the number of variables increases, so an exhaustive search is not a sensible approach to the problem [89]. Therefore, search strategies have been used to iterate comparisons on reduced sets of structures.

Scoring metrics [16, 92, 204, 379, 405] are developed to evaluate how well a particular Bayesian network structure fits with respect to a dataset and to guide the learning process. Classical goodness-of-fit criteria rank complex models higher than sparse ones. Nonetheless, a model should only have enough parameters to give an adequate representation of the association structure underlying the data. A criterion accounting for this tradeoff between model complexity and goodness-of-fit is the Bayesian information criterion (BIC) [405]. BIC provides a quantitative measure for model selection, penalizing the complexity of a model by an additional term which depends on the number of parameters and the sample size.

$$BIC = p(\mathcal{D}|\mathcal{S}, \boldsymbol{\theta}) - pen(\mathcal{N})\ dim(\mathcal{S}), \tag{3.6}$$

where $p(\mathcal{D}|\mathcal{S}, \boldsymbol{\theta})$ is the log-likelihood function, $pen(\mathcal{N})$ is equal to $log(N)/2$, being N the number of instances in the dataset, and $dim(\mathcal{S})$ is the network's dimension, that is, the number of independent parameters that have to be estimated.

The K2 scoring metric [98] computes the marginal likelihood of the dataset given the structure, subject to a uniform prior assumption on each variable data distribution. This scoring metric is decomposable, which facilitates the search process. Given the decomposability of the score, the marginal likelihood is maximized by maximizing, for each variable X_i, the expression:

$$g(X_i, \mathbf{\Pi}(X_i)) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!, \tag{3.7}$$

where r_i is the number of possible values of X_i; q_i is the number of possible values of $\mathbf{\Pi}(X_i)$; N_{ijk} is the number of cases in the database in which variable X_i takes its k-th value and $\mathbf{\Pi}(X_i)$ its j-th value; and N_{ij} is defined as $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

Search methods explore the space of Bayesian network structures and try to find a high scoring network structure. The most commonly used method is the greedy search algorithm [99] which uses a candidate network structure, which may be empty or provided by some expert, as a starting point. Then, at each iteration, this algorithm considers three possible operations: arc insertion, deletion or reversal. Next, the score is computed for every resulting candidate, and the candidate presenting the best score is selected and becomes the current candidate. This search process is iterated until there is no more score improvement.

The K2 algorithm [98] is one of the most famous Bayesian networks learning algorithms. This algorithm greedily learns a Bayesian network from a dataset by using the marginal likelihood score. Starting from the empty graph and a fixed order of the variables, this algorithm adds a variable as a parent to a given variable (from the subset of variables that are before this variable in the ordering only) whenever its inclusion represents an improvement in the marginal likelihood score. The algorithm stops the addition process when the marginal likelihood score decreases or the algorithm reaches the maximum admissible number of parents for each variable, which is fixed beforehand.

The search of the optimal Bayesian network structure have also tackled by means of other approaches. Some of these approaches are: simulated annealing [90], best-first search [261], estimation of distribution algorithms [46], ant colony optimization [116] and particle swarm optimization [104], among others. Despite this wide range of approaches, most researchers use genetics algorithms for the purpose of structure learning [142, 240, 276, 277, 311, 340, 441].

3.3.2 Parametric learning

Once the Bayesian network structure, $\mathcal{S}$, has been learnt, parameters, $\boldsymbol{\theta}$, have to be estimated from the dataset $\mathcal{D}$. Parameter learning aims to estimate the values of the conditional probability distribution of each variable X_i given any value of its parent set $\mathbf{\Pi}(X_i)$. Two well-known approaches for parametric learning are described in the following.

The maximum likelihood estimation (MLE) assesses the probabilities of variables from data without assuming any prior knowledge. It is based on the frequency of occurrences of variables in the data set, and selects the parameter configuration for a Bayesian network model, $\hat{\boldsymbol{\theta}}$, that maximizes the probability of the data set given the model $(\mathcal{S}, \boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \; p(\mathcal{D}|\mathcal{S}, \boldsymbol{\theta}) \; , \tag{3.8}$$

where $p(\mathcal{D}|\mathcal{S}, \boldsymbol{\theta})$ represents the likelihood function.

The maximum a posteriori (MAP) estimation is able to include prior knowledge into the parameter estimation problem. It selects the parameter configuration for a Bayesian network model, $\hat{\boldsymbol{\theta}}$, that maximizes the posterior probability of the parameters given the dataset $\mathcal{D}$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \; p(\boldsymbol{\theta}|\mathcal{D}), \tag{3.9}$$

3.4 Inference in Bayesian networks

Given the learned Bayesian network model, one of the most fundamental tasks for reasoning under uncertainty is evidence propagation [111, 281] which usually refers to computing the posterior probability of each single variable given the available evidence. In evidence propagation, a subset of variables $X_e \subset X$ (evidence variables) have been observed, and the goal is to reason about another subset of variables $X_q \subset X \setminus X_e$ (query variables) given the observed values of the evidence variables. This conditional probability can be computed as

$$p(X_q|X_e) = \frac{p(X_q, X_e)}{p(X_e)} \equiv \sum_{\boldsymbol{x_n}} \frac{p(X_n, X_q, X_e)}{p(X_e)}, \tag{3.10}$$

where $X_n \equiv (X \setminus X_e) \setminus X_q$ is the subset of non-observed variables.

Another interesting inference task is the total abduction problem. It consists of finding the most probable configuration of a set of variables of interest, X_q, given the evidence X_e. Similarly, the partial abduction deals with the problem of finding the values of only a subset of the query variables $X_q \subset X$ which yield the maximum posterior probability given the observed values for the evidence variables $X_e \subset X \setminus X_q$.

$$\hat{\boldsymbol{x}}_{\boldsymbol{q}} = \arg\max_{X_q} \; p(X_q|X_e) \; . \tag{3.11}$$

Although the process of probabilistic inference is proved to be NP-hard is some scenarios [97], the inference task is tractable for many real-world problems. The methods proposed for probabilistic inference can be divided into exact and approximate methods [192]. Many exact inference methods have been developed in the literature, such as [110, 255, 410, 429]. In contrast, approximate inference methods have been proposed when the previous exact inference methods are not suitable [78, 205, 236, 446].

3.5 Bayesian networks classifiers

The use of Bayesian network structures in supervised learning problems give rise to Bayesian classifiers [43, 163]. These classifiers model the joint probability distribution over the predictive variables $\boldsymbol{X}=\{X_1, X_2, \ldots, X_n\}$ and the class variable C. To classify an instance $\boldsymbol{x}$, the Bayes rule [37] is used to compute the posterior probability of each class label given the values of the predictive variables. The class with the maximum posterior probability is selected as the class label for the available instance. That is

$$\hat{c} = \arg\max_{c} p(c|\boldsymbol{x}) \propto \arg\max_{c} p(c)p(\boldsymbol{x}|c) \tag{3.12}$$

There exists a wide spectrum of Bayesian classifiers. A brief introduction of most important methods belonging to this classification approach is carried out.

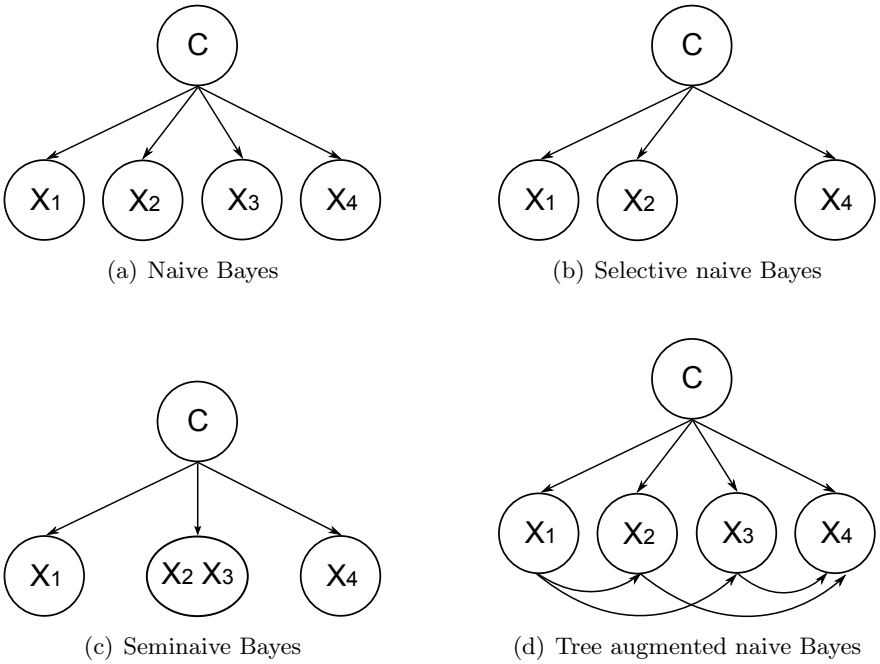
Naive Bayes Naive Bayes [325] is one of the simplest models for supervised classification. It is one of the most efficient and effective inductive learning algorithms for machine learning. Figure 3.3(a) represents the naive Bayes structure. The class variable, C, is discrete and takes values in the set $\Omega(C)$. The predictive features can be divided into two sets: the set of discrete features $\{X_1, \ldots, X_m\}$ and the set of continuous features $\{X_{m+1}, \ldots, X_n\}$. This classifier is based on Bayes theorem under the assumption of conditional independence of predictive features given the class variable. The naive Bayes classifier selects the most likely class value $\hat{c}$ of the posterior distribution:

$$\hat{c} = \arg\max_{c} \sum_{c=1}^{\Omega(c)} p(c) \prod_{i=1}^{m} p(x_i|c) \prod_{j=m+1}^{n} \mathcal{N}(x_j, \mu_j^c, (\sigma_j^c)^2) \ . \tag{3.13}$$

The above conditional independence assumption could be restrictive in several real-world problems, so, the following Bayesian classifiers are developed to alleviate this assumption.

Selective naive Bayes Selective naive Bayes [275] is a variant of naive Bayes that deals with correlated variables by selecting only a subset of the given variables into the final classifier. Figure 3.3(b) represents the selective naive Bayes structure where X_3 is excluded from the final model. This classifier improves accuracy in domains with redundant and irrelevant variables. The learning component adds the capability to exclude attributes that introduce dependencies to the original naive Bayes classifier. This greedy process consists of searching the space of attribute subsets. The direction of the search could be forward or backward. A forward selection method would start with the empty set and successively add variables, whereas a backward elimination process would begin with the full set and remove unwanted variables. The search process stops adding or eliminating attributes when none of the alternatives improves classification accuracy.

Seminaive Bayes Seminaive Bayes [260, 354] considers statistical relationships between variables in order to join them into new multidimensional ones that smooth the independence assumption of the naive Bayes. Figure 3.3(c) represents the seminaive Bayes structure where the new multidimensional variable is composed by the cartesian product of X_2 and



(a) Naive Bayes (b) Selective naive Bayes

(c) Seminaive Bayes (d) Tree augmented naive Bayes

Figure 3.3: Examples of Bayesian classifiers structures

X_3. Kononenko [260] relaxed the independence assumption by a restricted structure learning. His algorithm partitions the variables into disjoint groups and assumes independence only between variables of different groups. Pazzani [354] performed feature selection and feature joining to improve the naive Bayes. Specifically, he used forward sequential selections and joining (FSSJ) and backward sequential elimination and joining (BSEJ) to search dependencies and join the variables.

Tree augmented naive Bayes Tree augmented naive Bayes classifier [163] allows relationships between pairs of predictive variables in the network. It builds a dependence tree structure among the variables and, then, connect all the predictive variables with the class one. Figure 3.3(d) represents the tree augmented naive Bayes structure where the root node of the tree is X_1. Unlike the naive Bayes classifier, each predictive variable (except for the root node of the variable tree) has one additional predictive variable as a parent. Similarly, the k-dependence Bayesian classifier [391] also builds a dependence tree structure among the variables. In this case, it allows each predictive variable to have a maximum number of k parent variables. Both classifiers use the mutual information conditioned to the class variable to decide which edges are included and in which order. Its value is computed through the following expression:

$$I(X,Y|C) = \sum_{i=1}^{\Omega(X)} \sum_{j=1}^{\Omega(Y)} \sum_{k=1}^{\Omega(C)} p(x_i, y_j, c_k) log \frac{p(x_i, y_j|c_k)}{p(x_i|c_k)p(y_j|c_k)} \ , \tag{3.14}$$

where X and Y are discrete predictive variables conditioned to the class variable C.

Chapter 4

Scientometrics

4.1 Introduction

Scientometrics is the study of science, technology, and innovation from a quantitative perspective. This field has grown in popularity during last years and is used to describe the growth, structure, trend, interrelationship and productivity of science. Scientometrics represents the multiple facets of scientific activity in models of use to science policy makers, using quantitative tools with sound properties [491]. It uses the published works of scientists to answer the questions of policy makers, stakeholders and scientists themselves, among others, taking research and science as a research object. Major research issues include the measurement of impact, reference sets of articles to investigate the impact of journals and institutes, understanding of scientific citations, mapping scientific fields and the production of indicators for use in policy and management contexts [289].

Scientometrics has typically been defined as the "quantitative study of science and technology" [452]. Previously, Brusilovsky defined this field as the "study of the measurement of scientific and technological progress" [66], whereas Hess defined it as the "quantitative study of science, communication in science, and science policy" [206]. Other many definitions not covered here have been also proposed in the literature [65, 435, 465].

Modern scientometrics is mostly based on the work of Derek de Solla Price and Eugene Garfield. The 1960s and '70s saw the development of scientometrics as an operational activity for providing a response to the pressing demand for the measuring of science. The historian Derek de Solla Price published a number of books and articles which laid the foundations for the newly emerging field of quantitative science studies [118, 119, 120], culminating in a full-fledged research program [121]. In contrast, Eugene Garfield created the Science Citation Index [170] and founded the Institute for Scientific Information which is heavily used for scientometric analysis [95]. Other founding fathers of the discipline were Narin [343] in the United States of America, Nalimov and Mulchenko [342] in Russia and Braun and Bujdoso [59] in Hungary.

The origin of the term scientometrics goes back to the year 1969, when Nalimov and Mulchenko [342] coined the Russian equivalent of the term scientometrics ("naukometriya").

The term had gained wide recognition by the foundation in 1978 of the journal Scientometrics by Tibor Braun in Hungary. The launching of the journal Scientometrics that persuaded all those concerned that a self-contained research field under this name really exists. The journal became soon the leading information channel of the field.

Scientometrics is related to, and has similar interests with, bibliometrics, which is quantitative analysis of media in any written form. Although some works, like Campbell [77], Cole and Eales [94] and Hulme [217], are supposed to be the first bibliometric studies, the coining of the term bibliometrics is frequently credited to Pritchard [371] in 1969, who defined the new bibliometrics widely, to be "the application of mathematical and statistical methods to books and other media of communication". There are other many definitions of the term bibliometrics in the literature [71, 145, 200, 253, 254] which are not discussed here. There has been considerable confusion in the terminology of the two closely related metric terms [211]. The focus of bibliometrics has always been preponderantly on the literature per se of science, whereas scientometrics is not only focused on measuring the literature output (papers, books, patents, etc) but also on analyzing the practices of researchers, the socio-organizational structures, research and development management, the role of science and technology in the national economy, governmental policies towards science and technology, and so on.

Chapter outline

Section 4.2 introduces how citation analysis is used in scientometrics. The well-known *h-index*, its advantages and its drawbacks are presented in Section 4.3. According to the improvements of the *h-index*, Section 4.4 lists measures that complement the h-index, take time into account, allow for co-authorship and consider other variables. By journal measures, Section 4.5 presents the well-known impact factor and other measures that assess the quality of citations and correct for differences among fields. Finally, Section 4.6 introduces the main features of bibliographic databases like Web of Science, Scopus and Google Scholar.

4.2 Citation analysis in research evaluation

Citation analysis is one of the most widely used methods of scientometrics. This method uses citations in scientific works to establish links to other works or other researchers with the intention of analyzing the frequency, patterns, and graphs of citations in articles and books [174, 328]. Bibliometric measures have emerged from citation analysis to assess and compare the research activity of individual researchers according to their output. They constitute an objective method whose results are reproducible. The main advantage of these bibliometric measures is that they can summarize the scientific production of a researcher as a set of quantitative figures that permit rapid comparison. This can at the same time be a limitation, because it removes many details from the citation records. Nowadays, many funding agencies and promotion committees use bibliometric measures regularly as a decision-support tool to evaluate the impact of research projects and researchers alike. Also, bibliometric measures

are commonly adopted for the purpose of allocating public funds. These measures essentially involve counting the number of times scientific papers are cited. They are based on the assumption that influential researchers and important papers will be cited more frequently than others. Thus, bibliometric measures are an increasingly important topic within the scientific community.

The process of evaluation of scientific research has become a central, difficult and lengthy process in the management and governance policies of national research systems [283]. Almost every research assessment decision (accepting research projects, contracting researchers, awarding scientific prices, concede a grant and so on) depends to a great extent upon the scientific merits of the involved researchers. The most well-known method used for researchers assessment is peer review. This process involves some reviewers reading and discussing scientists' papers to determine the validity of the ideas and results, and their potential impact on the world of science. Peer assessment is undoubtedly the principal procedure for judging quality. However, it has been mooted that peer assessment and similar expert-based evaluations have serious shortcomings and disadvantages [96, 450]. The opinions of experts are linked to subjectivity and may have conflict of interest elements or be the result of unrelated factors and negative or positive biases. Although if used properly peer review is assumed to be the most reliable system, it is slow, expensive and unwieldy [93, 338, 393]. Other authors contest this appraisal [27, 195, 212]. This difference of opinion among authors has led to the development of bibliometric measures as a method for researchers assessment. These measures may contribute to the fairness of research evaluations by presenting objective and impartial information to a peer review that would otherwise depend more on the personal views and experiences of the scientists appointed as referees. Both types of methods have pros and cons, extensively discussed in the literature [213, 304, 334, 453], in terms of costs, execution times, limitations and objectiveness. It appears that these methods can coexist, but not always in an easy and synchronized fashion. Sometimes it appears that the two methods of evaluating research quality tend to contradict or oppose each other [245].

Several bibliometric measures have been developed in the literature (see reviews [17, 138]. An obvious measure is citation count, which quantifies the impact of the cited work [170, 173]. In spite of its simplicity, the main recognized disadvantages of this standard bibliometric measure is that it does not reflect the full impact of the scientific research and is extremely affected by a single highly-cited paper [353]. Thus, the average citation rate [403] was proposed as total quality of the research output. Beyond these traditional measures, one of the most successful bibliometric measures was proposed by Jorge Hirsch and it was called the h-index [207]. It quantifies the scientific output of a single researcher as a single-number criterion. It is a simple new measure incorporating both the quantity and visibility of publications. Since its introduction, the *h-index* has received a lot of attention from other researchers. In the Web of Science Hirsch's article has been cited more than 1,800 times (November, 2014).

4.3 The h-index

The *h-index* [207] is a single measure that combines papers (an aspect of quantity) and citations (an aspect of quality) to characterize researchers' scientific output. Considering a researcher's list of publications, ranked according to the number of citations received, the *h-index* is defined as the highest rank such that the first h publications received each at least h citations. Formally, Hirsch defined the *h-index* as follows: "A scientist has index h, if h of his or her N_p papers have at least h citations each and the other (N_p - h) papers have no more than h citations each". According to the example of Table 4.1, it is observed that the *h-index* value is 6. The papers on ranks 1,...,h constitute the so-called *h-core* [384].

Although Hirsch defined the *h-index* in 2005, it appears that the *h-index* (of course not with this name) was defined some 35 years earlier by Arthur Stanley Eddington in a communication of Eddington to the Harold Jeffreys [131]. In order to record his cycling prowess, Eddington records n, being the highest number of days on which he had cycled n or more miles. This is nothing else than Hirsch's index but in cycling terminology.

A short paper published in Nature made the *h-index* known to many scientists [28]. Undoubtedly, the *h-index* synthetically aggregates two important aspects of the scientist's production: impact, represented by the number of citations per paper, and productivity, represented by the number of different papers. Hirsch originally suggested the *h-index* for application at the micro level, that is, as a measure to quantify the scientific output of a single researcher. It has been used at an even further micro level to assess single highly cited publications [401]. However, the *h-index* can be also used at the meso and macro levels by means of *successive h-indices* [367, 403]. Thus, the *h-index* can be also applied to journals [60, 403], research groups [454], institutions [332, 367], publishers [400] and countries [108]. Finally, the *h-index* has not been only used for scientific comparisons but also for detecting interesting hot topics and compounds in diverse research areas [30], and for predicting future achievements [208], among others.

Table 4.1: Calculating bibliometric measures of a fictitious researcher

Features	Papers										
$rank$	1	2	3	4	5	6	7	8	9	10	11
$cits$	28	17	12	10	9	6	2	2	2	1	0
$rank^2$	1	4	9	16	25	36	49	64	81	100	121
$\sum cits$	28	45	57	67	76	82	84	86	88	89	89
$year$	2009	2008	2010	2010	2012	2011	2012	2013	2014	2013	2014
age	6	7	5	5	3	4	3	2	1	2	1
$cits/age$	4.6	2.4	2.4	2.0	3.0	1.5	0.6	1.0	2.0	0.5	0.0
$authors$	4	3	2	2	3	3	3	4	1	2	3
$cits/authors$	7.0	3.4	6.0	5.0	3.0	2.0	0.6	0.5	2.0	0.5	0.0
r_{eff}	0.25	0.58	1.08	1.58	1.91	2.25	2.58	2.83	3.83	4.33	4.66
$position$	1st	2nd	1st	1st	2nd	1st	2nd	3rd	1st	1st	2nd
$S(a,d)$	0.40	0.33	0.66	0.66	0.33	0.50	0.33	0.20	1.00	0.66	0.33

In his original proposal, Hirsch pointed out good properties of the *h-index*. For example, he stated that the proposed *h-index* measures the broad impact of an individual's work, avoids all of the disadvantages of traditional criteria (total number of papers, total number of citations, citations per paper, number of significant papers, etc.), is easy to compute by ordering papers by "times cited" field in the Web of Science, among others.

Other works like Costas and Bordons [100] and Glanzel [180] and Vanclay [456] also pointed out the advantages of the *h-index*. Some of the most interesting properties of the *h-index* are summarized in the following:

- It combines citation impact with publication activity measures.
- It performs better than other single criteria commonly used in research evaluation.
- It can be applied to any level of aggregation but is used at individual scientists level.
- It measures durable performance, not only single peaks.
- It is not immediately affected by increasing publications alone.
- It is insensitive to a set of lowly cited papers.
- It is an objective indicator which may play an important role when making decisions about promotions, fund allocation and awarding prizes.

However, the *h-index* presents some disadvantages that have been pointed out in the literature [51, 136, 227, 251, 252, 310, 395, 448, 464, 476]. Different authors have tried to overcome those drawbacks by defining new indicators, see Section 4.4. Hirsch himself noted that there are inter-field differences in typical *h-index* values due to differences among fields in productivity and citation practices, so the *h-index* should not be used to compare scientists from different disciplines [207]. Moreover, Hirsch noted that there exist some technical limitations, such as the difficulty to obtain the complete output of scientists with very common names. It is important to remark that these drawbacks are shared with almost any indicator that is based in citation counts. Also, the *h-index* depends on the duration of each scientist's career because the pool of publications and citations increases over time. In order to overcome this limitation, Hirsch presented the "m parameter", which is the result of dividing the *h-index* by the scientific age of a scientist (number of years since the author's first publication). Other shortcomings of the *h-index* are presented in the following:

- It is insensitive to highly cited papers, being the excess citations completely ignored once papers are included into the *h-core*.
- It does not show decay in a scientist's carrier since the number of citation might increase even if no new papers are published.
- It does not take into account in any way the number of coauthors of each paper.

- It does not take into account the distribution of the citations in the *h-core*.
- Due to its simple computation, there is a risk of indiscriminate use, such as relying only on it for the assessment of scientists.
- Its use could provoke changes in the publishing behavior of scientists, such an artificial increase in the number of self-citations.
- It is suited for the micro level but at higher levels of aggregation there are more versatile indicators.
- The application of appropriate indicators sets instead of one single measure can provide a more adequate and multifaceted picture of reality.

4.4 Improvements of the h-index

Despite of the good properties of the *h-index*, many authors have pointed out several drawbacks of the indicator (see Section 4.3). To overcome these drawbacks many new bibliometric measures have been proposed in the literature. For clarity reasons, these measures are categorized into four groups: (i) measures that try to complement the *h-index*, (ii) measures that extend the *h-index* to take time into account, (iii) measures which analyze how to count multi-authored publications, and iv) measures which take into account other variables. Their main properties are summarized in the following.

4.4.1 Bibliometric measures that complement the h-index

The *h-index* has been extended by many authors that have proposed new variations of the *h-index* that try to overcome its main drawbacks. For example, the *g-index* [134], the *e-index* [489] and the *tapered h-index* [20] are proposed to take citations that are completely ignored by the *h-index* calculation into account.

g-index Since the *h-index* tends to underestimate the achievement of researchers that have a selective publication strategy, that is, researchers that do not publish a lot of documents but have a major international impact, the *g-index* [134, 135, 136] was proposed as being sensitive to the level of the highly cited papers. It is defined as the highest rank such that the cumulative sum of the number of citations received is greater than or equal to the square of this rank. According to Table 4.1, the *g-index* value is 9. It is the highest rank such that the top 9 papers have at least 9^2=81 citations (here $88 > 81$); on rank 10 we have $89 < 10^2 = 100$ citations. Unlike the *h-index*, the *g-index* takes into account the exact number of citations received by highly cited papers, favoring researchers with a selective publication strategy. A criticism of the *g-index* was raised observing that the highest rank could be larger than the total number of author's publications. However, the greatest drawback of the *g-index* is that it may be greatly influenced by a unique very successful paper. Finally, although the *g-index* is better than the *h-index* in this sense, is not a fully satisfactory solution.

e-index Although effective and simple, the *h-index* suffers from some drawbacks that limit its use in accurately and fairly comparing the scientific output of different researchers. Aimed to the same goal as the *g-index*, the *e-index* [489] was proposed to represent the excess citations that are completely ignored by the *h-index* calculation. As a mathematical formula the *e-index* is defined as

$$e\text{-}index = \sum_{i=1}^{h}(cit_i - h) \tag{4.1}$$

where h is the value of *h-index*, and cit_i is the number of citations of paper i. According to Table 4.1, the *e-index* value is (28-6)+(17-6)+(12-6)+(10-6)+(9-6)+(6-6) = 46.

tapered h-index The main idea of the *tapered h-index (h_t-index)* [20] is to take into account all the citations, giving to each of them a value equal to the inverse of the increment that would suppose to increase *h-index* one unit. It is defined as

$$h_t\text{-}index = \sum_{i=1}^{N_p} h_t(i), \text{ where } h_t(i) = \begin{cases} \dfrac{cit_i}{2i-1} & \text{if} \quad cit_i \leq i \\ \dfrac{i}{2i-1} + \displaystyle\sum_{j=i+1}^{cit_i} \dfrac{1}{2j-1} & \text{if} \quad cit_i > i \end{cases} \tag{4.2}$$

where N_p is the number of total publications and cit_i is the number of citations of paper i. According to the values of Table 4.1, the h_t-*index* can be calculated as

i	cit_i	$h_t(i)$	i	cit_i	$h_t(i)$
1	28	$\frac{1}{1} + \sum_{j=2}^{28} \frac{1}{2j-1} = 2.64$	6	6	$\frac{6}{11} = 0.54$
2	17	$\frac{2}{3} + \sum_{j=3}^{17} \frac{1}{2j-1} = 1.73$	7	2	$\frac{2}{13} = 0.15$
3	12	$\frac{3}{5} + \sum_{j=4}^{12} \frac{1}{2j-1} = 1.29$	8	2	$\frac{2}{15} = 0.13$
4	10	$\frac{4}{7} + \sum_{j=5}^{10} \frac{1}{2j-1} = 1.02$	9	2	$\frac{2}{17} = 0.11$
5	9	$\frac{5}{9} + \sum_{j=6}^{9} \frac{1}{2j-1} = 0.84$	10	1	$\frac{1}{19} = 0.05$
			11	0	$\frac{0}{21} = 0.00$

where the h_t-*index* is 2.64+1.73+1.29+1.02+0.84+0.54+0.15+0.13+0.11+0.05+0.00=8.55.

Other indices like the *a-index* [242], the *m-index* [55] and the *r-index* [243] are developed to measure the citation intensity of the *h-core* papers.

a-index The average number of citations received by the articles included in the *h-core* is represented by the *a-index* [242]. This index measures the citation intensity of the *h-core* papers; however, it can be very sensitive to just a few papers receiving high citation counts. It achieves the same goal as the *g-index*, namely correcting for the fact that the original *h-index* does not take the exact number of citations of articles included in the *h-core* into account.

Mathematically, it is defined as:

$$a\text{-}index = \frac{1}{h}\sum_{i=1}^{h} cit_i \tag{4.3}$$

where h is the *h-index* value and cit_i is the numbers of citations of paper i. According to Table 4.1, it is observed that the *a-index* value is (28+17+12+10+9+6)/6 = 13.66.

m-index As the distribution of citation counts is usually skewed, the median and not the arithmetic mean should be used as the measure of central tendency. Therefore, a new index, called *m-index* [55], is proposed as a variation on the *a-index*. The *m-index*, which was designed to illustrate the impact of the papers in the *h-core*, is the median number of citations received by the h most visible papers. According to Table 4.1, it is observed that the *m-index* value is 11.

r-index In order to overcome some problems related to the *a-index*, a new measure, called *r-index* [243], is proposed. Unlike the *a-index*, which involves a division by the *h-index*, the *r-index* does not punish researchers for having a higher *h-index* value. Therefore, instead of dividing by the *h-index*, the *r-index* takes the square root of the sum of citations in the *h-core* to calculate the citation intensity of the h most visible papers. Like the *a-index*, the *r-index* can be also very sensitive to just a very few papers receiving extremely high citation counts. As a mathematical formula the *r-index* is defined as:

$$r\text{-}index = \sqrt{\sum_{i=1}^{h} cit_i} \tag{4.4}$$

where h is the *h-index* value and cit_i is the numbers of citations of paper i. According to Table 4.1, it is observed that the *r-index* value is $\sqrt{28+17+12+10+9+6} = 9.06$. Finally, the *r-index* can be also computed as $\sqrt{a \cdot h}$, where a and h are the *h-* and *a-index* values.

This section also introduces some measures which are useful to distinguish individuals with the same *h-index* value. The *rational h-index* [386], the *multidimensional h-index* [168] and the *individual annual h-index* [197] are described in the following.

rational h-index As an extension of the original *h-index*, the *rational h-index* [386] is proposed to take into account the number of citations needed to increase the *h-index* by one unit. It measures the distance to the next value of the *h-index*. Mathematically, this is

$$rational\ h\text{-}index = h + 1 - \frac{n}{2h+1} \tag{4.5}$$

where h is the value of the *h-index*, and n is the number of citations needed to increase the *h-index* by one unit. In this context, the *rational h-index* value can be calculated as

$$\textit{rational h-index} = 6 + 1 - \frac{6}{13} = 6.53.$$

multidimensional h-index For the purpose of distinguishing among individuals with the same *h-index* value, the *multidimensional h-index* [168] is proposed. It uses the same logic as the original *h-index* and provides additional information under the same principles. The *multidimensional h-index* is composed by a set of components in which the conventional *h-index* value is only the first component. Additional components of the *multidimensional index* are obtained by computing the *h-index* for the subset of papers not considered in the immediately preceding component. This process iterates to obtain subsequent *h-index* values until all cited papers are analyzed. According to Table 4.1, the *multidimensional h-index* is formed by the set (6, 2, 1, 1). This extension is useful in fields where *h-index* values are generally low.

i	cit_i	i	cit_i	i	cit_i	i	cit_i
1	**28**						
2	**17**						
3	**12**						
4	**10**						
5	**9**						
6	**6**						
7	2	**1**	**2**				
8	2	**2**	**2**				
9	2	3	2	**1**	**2**		
10	1	4	1	2	1	**1**	**1**
11	0	5	0	3	0	2	0
1^{st}*h-index*: **6**		2^{nd}*h-index*: **2**		3^{rd}*h-index*: **1**		4^{th}*h-index* **1**	

individual annual h-index This index [197] is proposed to represent the average annual increase in the academic's individual *h-index*. It provides a more reliable comparison between academics in different disciplines and at different career stages than the original *h-index*.

In order to provide a more balanced view of scientific production some measures like the *hg-index* [18] and the q^2*-index* [74] are proposed.

hg-index With the intention of keeping the advantages of both *h-index* and *g-index* as well as to minimize their disadvantages, the *hg-index* [18] is developed. Both measures incorporate several interesting properties about the publications of a researcher and that both should be taken into account to measure the scientific output of scientists. Therefore, the *hg-index* is

a combined index which characterizes the scientific output of researchers. It is computed as the geometric mean of the *h-index* and *g-index*:

$$hg\text{-}index = \sqrt{h \cdot g} \tag{4.6}$$

where h is the *h-index* value and g is the *g-index* value. According to Table 4.1, it is observed that the *hg-index* value is $\sqrt{6 \cdot 9} = 7.35$.

q²-index In order to provide a more global view of scientific production, the *q^2-index* [74] is developed. It is based on the geometric mean of the *h-index*, describing the number of the papers (quantitative dimension), and the *m-index*, depicting the impact of the papers (qualitative dimension). As a mathematical formula the *q^2-index* is defined as:

$$q^2\text{-}index = \sqrt{h \cdot m} \tag{4.7}$$

where h is the *h-index* value and m is the *m-index* value. According to the example of Table 4.1, the *q^2-index* value is $\sqrt{6 \cdot 11} = 8.12$.

Finally, other indices, which are defined in a similar way to the *h-index*, are also proposed. The *h(2)-index* [262], the *wu-index* [480] and the *h_α-index* [445] are examples belonging to this category.

h(2)-index A scientist's *h(2)-index* [262] is defined as the highest natural number such that his *h(2)* most-cited papers received each at least *$h(2)^2$* citations. The advantage of *h(2)-index* as the index to characterize the scientific output of an individual over the original *h-index* is that less work is required to verify the authorship of the relevant papers. In this context, it is observed that the *h(2)-index* value is 3. It is the highest rank such that the top 3 papers have each at least 3^2=9 citations (here $12 > 9$).

wu-index It assesses the substantial impact of a researcher's work. The *wu-index* [480] is defined in a similar way to the *h-index* but focusing only in excellent papers (or highly cited papers). To do so, it is expressed as: If wu of a researcher's papers have at least $10wu$ citations each and the other papers have fewer than $10(wu + 1)$ citations, that researcher's *wu-index* is wu. Using the above definition, the *wu-index* value is 1 according to Table 4.1. It is the highest rank such that the top 1 papers have each at least $10 * 1$ citations.

h_α-index A new measure of scientific performance is proposed to generalize the *h-index*. This new measure depends on a parameter α and is therefore referred to as the *h_α-index* [445]. It is expressed as: A researcher has *h_α-index* h_α, if h_α of his/her articles received at least $h_\alpha \cdot \alpha$ citations each, and the rest articles have received no more than $h_\alpha \cdot \alpha$ citations. Clearly, for α=1 the *h_α-index* reduces to the *h-index*. Furthermore, for α=10 the *h_α-index* reduces to the *wu-index*.

4.4.2 Bibliometric measures that take time into account

The original *h-index* does not take into account the age of an article. It may be the case that some scientist contributed a number of significant articles that produced a large *h-index*, but now he/she is rather inactive or retired. Therefore, senior scientists, who keep contributing nowadays, or brilliant young scientists, who are expected to contribute a large number of significant works in the near future but now they have only a small number of important articles due to the time constraint, are not distinguished by the original *h-index*. Thus, it arises the need of defining some measures in order to account for these facts, among others.

ar-index As an adaptation of the *r-index*, the *ar-index* [244] is proposed to take into account not only the citation intensity in the Hirsch core but also makes use of the age of the publications in the *h-core*. Therefore, the *ar-index* not only can increase but also decrease over time. It is defined as the square root of the sum of the average number of citations per year of articles included in the *h-core*. Mathematically, it is defined as:

$$ar\text{-}index = \sqrt{\sum_{i=1}^{h} \frac{cit_i}{age_i}} \tag{4.8}$$

where h is the *h-index* value, cit_i is the numbers of citations of paper i, and age_i is the age of paper i. Table 4.1 shows the *ar-index* value is $\sqrt{4.6+2.4+2.4+2.0+3.0+1.5}$=3.99.

contemporary h-index The original *h-index* cannot distinguish between inactive scientists, young scientists and senior scientists who are still contributing nowadays. For this reason, a *contemporary h-index* [415] is defined to take into account the age of papers. In this way, the value of old papers gradually declines, even if they still receive citations. The *contemporary h-index* is expressed as follows: "A researcher has *contemporary h-index* h^c if h^c of its N_p articles get a score of $S_i^c \geq h^c$ each, and the rest (N_p - h^c) articles get a score of $S_i^c \leq h^c$", where S_i^c is defined as

$$S_i^c = \gamma \cdot age_i^{-\delta} \cdot cit_i \tag{4.9}$$

where age_i is the age of paper i, cit_i is the number of citations received by paper i, and γ and δ are arbitrary parameters. Fixing γ and δ equal to 1, the above score S_i^c is calculated as cit_i/age_i. Therefore, ranking S_i^c in descending order, the *contemporary h-index* value is 2 (see Table 4.1) because it is the highest rank such that the top 2 papers have at least 2 citations per year.

.

trend h-index The *h-index* does not take into account the year when an article acquired a particular citation, that is, the age of each citation. It cannot identify scientists whose work is considered pioneering and sets out a new line of research that currently is hot. Thus, a

trend h-index [415] is defined for the above purpose. Unlike the *contemporary h-index*, the *trend h-index* assign to each citation of an article an exponentially decaying weight, which is expressed as a function of the age of the citation. It is expressed as follows: "A researcher has *trend h-index* h^t if h^t of its N_p articles get a score of $S_i^t \geq h^t$ each, and the rest (N_p - h^t) articles get a score of $S_i^t \leq h^t$", where S_i^t is defined as

$$S_i^t = \gamma \cdot \sum_{\forall x \in cit_i} age_x^{-\delta} \tag{4.10}$$

where cit_i is the number of citations received by paper i, age_x is the age of citation x, and γ and δ are arbitrary parameters.

age decaying h-index In order to take into account both the age of the scientific's article and the age of each citation to the article, the *age decaying h-index* [247] is proposed. It is a generalization of both the *contemporary h-index* and *trend h-index* and it is expressed as follows: "A researcher has *age decaying h-index* h_{ad} if h_{ad} of its N_p articles get a score of $S_i^{ad} \geq h_{ad}$ each, and the rest (N_p - h_{ad}) articles get a score of $S_i^{ad} \leq h_{ad}$", where S_i^{ad} is defined as

$$S_i^{ad} = \gamma^2 \cdot age_i^{-\delta_1} \cdot \sum_{\forall x \in cit_i} age_x^{-\delta_2} \tag{4.11}$$

where age_i is the age of paper i, cit_i is the number of citations received by paper i, age_x is the age of citation x, and γ, δ_1, and δ_2 are arbitrary parameters.

f-index As remarked before, the *h-index* does not take into account the time-width of the scientific research. Some variants like the *contemporary h-index* and the *age decaying h-index*, among others, give less importance to older articles introducing an age-related weighting system. However, these attempts complicate the calculation of a final synthetic indicator, introduce some weights and parameters which can be questionable. In contrast, the *f-index* [160] takes into account the age of the publications in a simpler way. It computes the time-range of the papers with at least one citation (added to 1 to consider the time spent to prepare the first paper included in the set):

$$\textit{f-index} = range_{i \in \Omega}\,(y_1, y_2, \ldots, y_i, \ldots) + 1 \tag{4.12}$$

where y_i is the publication year related to paper i and Ω is the set of publications which have been cited at least once. According to Table 4.1, the *f-index* value is *range* (2009, 2008, 2010, 2010, 2012, 2011, 2012, 2013, 2014) + 1 = 7. One of its main characteristics is that it does not compromise the original simplicity and immediacy of understanding of the *h-index*.

hpd-index With the intention of comparing the scientific output of researchers in different ages, the *hpd-index* [263] is developed. To do so, the average number of citations per decade is defined as a measure of success of a scientific paper. The *hdp-index* is defined in the

following way: "A scientist has index *hpd* if *hpd* of his/her papers have at least *hpd* citations per decade each, and his/her other papers have less than *hpd* + 1 citations per decade each".

h-index sequences Let the career period of a researcher be described by time t=$t_1, t_2, ..., t_n$, where t_1 denotes the year of the first publication and so on, until t_n, the final year of the career or the present year. Then, the *h-index sequence* [290], which is proposed to quantify the progress of scientists' careers, is constructed as follows. If it is only considered the papers of publication year t_n and their citations obtained in the same year, then the first *h-index* of the sequence, denoted as h_1, can be derived. Next, the years t_n and t_{n-1} together and their citations obtained in the same period, yielding the next *h-index*, denoted as h_2. The process continues until the t_1 year is reached. The last *h-index* of the sequence, denoted as h_n, is obtained considering all years $t_1, \ldots, t_n$ and taking into account all publications and citations to these publications in this period. Finally, the *h-index sequence* $h_1, h_2, \ldots, h_n$ gives a dynamic description of the visibility of the researcher's career. While the original *h-index sequence* is calculated using time in the reverse way (in the direction of the past), it seems more logical to use the time in forward direction (in the direction of the present). The above approach [70, 137] leads to real career *h-index sequences*.

4.4.3 Bibliometric measures that allow for co-authorship

The problem of how to count multi-authored publications has been discussed for a long time [293]. Several methods for accrediting publications for several authors have been developed in the literature [95, 122, 123, 139, 447, 493]. A short overview of some scoring methods follows.

- Total counting: Each of the n authors receives one credit. This counting method is also called normal, or standard counting.
- First-author counting: Only the first of the n authors of a paper receives a credit equal to one. The other authors do not receive any credit.
- Fractional counting: Each of the n authors receives a score equal to $1/n$. This counting method is sometimes called adjusted counting.
- Proportional counting: If an author has rank r in the author list of an article with n collaborators, then the author receives a score of n+1-r. This score can be normalized in such a way that the total score of all authors is equal to 1. In this normalized version the score is: $\frac{2}{n}(1 - \frac{r}{n+1})$.
- Geometric counting: If an author has rank r in an article with n collaborators, then the author receives a credit of 2^{n-r}. In its normalized version this score becomes $\frac{2^{n-r}}{2^n-1}$.
- Noblesse oblige: In this approach it is assumed that the most important author closes the list. This author receives a credit of 0.5, while the other $n-1$ authors receive a credit of $\frac{1}{2(n-1)}$ each.

It has been shown that the number of citations a paper receives can be influenced by the number of authors since the greater the number of authors, the greater the number of self-citations [181]. Perhaps the most important shortcoming of the *h-index* is that it does not take into account in any way the number of coauthors of each paper. In the following, some well-known indices which account for the co-authorship effect is detailed. Other indices not covered in this section are: the *golden productivity index* [24], the *adapted pure h-index* [82], the $\bar{h}$*-index* [209], the *fractional p-index* [369] and the g_m*-index* [399], among others.

h_i-index Since the *h-index* can be inflated if a scientist has written many co-authored articles, the h_i*-index* [36] is proposed to consider the idea of taking collaboration into account. It estimates the number of papers that a researcher would have written throughout his or her career with at least h_i citations if he or she had worked alone. The rationale behind is to measure the effective individual average productivity. The h_i*-index* simply divides the *h-index* value by the average number of researchers in the publications of the *h-core*. Mathematically, it is defined as

$$h_i\text{-}index = \frac{h}{N_a} = \frac{h^2}{N_a^T} \tag{4.13}$$

where h is the value of the *h-index*, and N_a is the mean number of authors in the *h-core*, and N_a^T is the total number of authors in the *h-core*. According to Table 4.1, the h_i*-index* value is $6^2/(4+3+2+2+3+3) = 2.11$. Authors used the mean number of authors of the papers in the *h-core* as the factor with which to rationalize the *h-index* and obtained a fractional value that accounts for multiple authorship. The average is sensitive to extreme values and therefore the normalization with the mean number of authors disfavors people with some papers with a large number of co-authors. As an alternative they propose dividing by the median number of researchers.

h_f-index This index takes into account the co-authorship effect dividing the number of citations by the number of authors for each paper. The h_f*-index* [397] is expressed as follows: A researcher has h_f*-index* h_f if h_f of its articles get a ratio at least equal to h_f. However, this has the disadvantage that for a determination of the h_f*-index* the publications have to be rearranged into a new order according to this ratio. According to Table 4.1, the h_f*-index* value is 3 because it is the highest rank such that the top 3 papers have each at least 3 citations per author. This index leads to the strange effect that highly cited papers may not contribute to the index because they have a large number of authors, so that they drop out of the core by the rearrangement.

h_m-index A new index, called the h_m*-index* [398], pretends to soften the influence of the number of co-authors for a researcher's publications. To do so, it counts the papers fractionally according to the number of authors. This yields an effective number r_{eff} which is utilized to define the h_m*-index* as that effective number of papers that have been cited h_m or more times. Mathematically, it is defined as

$$h_m\text{-}index = max_r(r_{eff}(r) \leq cit(r)) \tag{4.14}$$

where $r_{eff}(r) = \sum_{i=1}^{r} 1/a_i$, a_i is the number of authors of paper i, and $cit(r)$ is the number of citations of paper r. According to Table 4.1, it is observed that the h_m-*index* value is 6. It is the highest rank such that the 6th paper has at least r_{eff} citations (here $6 > 2.25$), and on rank 7 we have $2 < 2.58$.

pure h-index This index takes the actual number of co-authors and the relative position of an author in the byline into account. A normalized score like first-author counting, proportional counting, and geometric counting, among other, is needed to be defined to compute the *pure h-index* [470]:

$$pure\ h\text{-}index = h \cdot \sqrt{\frac{h}{\sum_{d=1}^{h} S(a,d)^{-1}}} \tag{4.15}$$

where h is the *h-index* value of author a, and *S(a,d)* denotes the normalized score of author a in document d. The term normalized refers to the fact that the sum of all scores of one document must be one. Fixing proportional counting as an example of normalized score, the h_p-*index* value is

$$pure\ h\text{-}index = 6 \cdot \sqrt{\frac{6}{0.40^{-1} + 0.33^{-1} + 0.66^{-1} + 0.66^{-1} + 0.33^{-1} + 0.50^{-1}}} = 3.98$$

4.4.4 Bibliometric measures that consider other variables

This section reviews some measures which consider other aspects like the number of references, the number of different citers, the citation speed, the field normalization, and citation distribution, among others.

creativity index This index is based on the creation of new scientific knowledge. The *creativity index* [423] tries to highlight papers that receives many citations and have few references. After splitting the creativity of each paper among its authors, the cumulative creativity of an author is then proposed as an indicator of her or his merit of research. Mathematically, it is defined as

$$c\text{-}index = \sum_{i=1}^{N_p} \frac{c(n_i, m_i)}{a_i}, \tag{4.16}$$

where $c(n_i, m_i) \simeq m_i - n_i + \frac{n_i}{Ae^{az} + Be^{bz}}$, N_p is the total number of papers; n_i is the number of references of paper i; m_i is the number of citations of paper i; a_i is the number of authors of paper i; $z = (m_i - 1)/(n_i + 5)$; and A, B, a, b are arbitrary parameters.

ch-index The proposed *ch-index* [15] is defined as the number such that, for a general group of scientific publications, *ch* publications are cited by at least *ch* different citers, while the other publications are cited by no more than *ch* different citers. According to the definition of the *ch-index*, if the same citing author cites a publication more than one time, then the author has to be counted only once. The most important benefit of the *ch-index* is to be insensitive to self-citations and citations made by recurrent citers.

citation speed index The scientific impact of a publication can be determined not only based on the number of times it is cited but also based on the citation speed with which its content is noted by the scientific community. In this context, the *citation speed index* [53] is proposed. Its calculation is based on the number of months that have elapsed since the first citation. The *citation speed index* is defined as follows: a group of papers has the *citation speed index* s if for s of its N_p papers the first citation was at least s months ago, and for the other (N_p - s) papers the first citation was $\leq s$ months ago.

success-index The *success-index* [159] is defined as the number of papers with a number of citations greater than or equal to *cit*. This number *cit* is an estimate of the number of citations that a publication should potentially achieve in a certain scientific context and period of time. The most complicated operation when constructing the *success-index* is determining the *cit* value of each paper. To do this, there are different possible approaches, which are borrowed from the existing literature on field-normalized indicators. Because of the fact that success status of a specific paper is determined independently on the other papers of interest, the *success-index* can be applied to groups of papers from different disciplines.

percentile-based indicator The *b-index* [56] is defined as the number of papers in the publication set of a scientist that in the individual publication years belong to the top 10% of most cited papers in a field. In contrast, the *x-index* [380] is formulated to estimate the level of research excellence and analyzes the numbers of papers in the top 1% and 0.1% of highly cited papers.

w-index Unlike the *h-index* that tends to cluster many researchers into the same index value, the *w-index* [479] leads to a somewhat finer ranking because its range could be up to twice the range of the *h-index*. The *w-index* is expressed as follows: "A *w-index* of at least k means that there are k distinct publications that have at least $1, 2, \ldots, k$ citations, respectively".

4.5 Journal-based measures

Citation analysis is one of the most widely used bibliometric tools for ranking journals. Garfield proposed the *impact factor*, which is a citation-based measure, as a fundamental tool in journal evaluation [171]. It was the first journal citation index to be calculated for

a large set of journals. It is calculated as the number of citations a journal receives in a given year to items published in the previous two years divided by the number of articles published in the previous two years. Although the *impact factor* has been widely used to assess journal performance, it discards much of the useful information that is present in the full citation network. For example, the *impact factor* does not account for where citations come from, that is, citations from prestigious journals are worth no more than citations from lower-tier publications. The advantages and drawbacks of the *impact factor* are presented in the following. Also, other journal citation measure, which are proposed to overcome some limitations of the *impact factor*, are reviewed in this section.

4.5.1 Impact factor

Without any doubt, the *impact factor* [170] is the most prominent citation measure to evaluate the relative influence, importance or prestige of scholarly journals. It has assumed so much power since it is starting to control the scientific enterprise and plays a crucial role in hiring, tenure decisions, and the awarding of grants. According to the Journal Citation Reports, the *impact factor* "is basically a ratio between citations and citable items published. Thus, the 2014 impact factor of journal X would be calculated by dividing the number of all the source journals' 2014 citations of articles journal X published in 2012 and 2013 by the total number of citable source items it published in 2012 and 2013". Thus, the *impact factor* is "a measure of the frequency with which the average cited article in a journal has been cited in a particular year" [171]. Mathematically, it is defined as

$$\textit{impact factor}\ (v_i, t) = \frac{\sum_j c(v_j, v_i, t)}{n(v_i)} \tag{4.17}$$

where $c(v_j, v_i, t)$ corresponds to the number of citations from journal v_j to journal v_i in year t. The number of publications published in journal v_i, denoted $n(v_i)$, during the two years previous to t, normalizes the resulting citation count, leading to a mean, 2-year citation rate per article.

The numerator of the *impact factor* considers the journal as a whole, and includes any citations to the journal title [41], thus it depends on accurate and complete aggregation of citations to a journal title. The denominator considers the journal as a collection of items that are likely to influence the scholarly literature, by way of citation; thus only the citable items in the journal are included. The composition of the denominator [319] is based on analysis of the content of the journal and the bibliographic parameters of its published source items.

The Journal Citation Reports also regularly publishes other specific journal citation indicators like the *immediacy index* and the *cited half-life* of journals [305]. The *immediacy index* provides the number of citations an item obtains in the year of publication itself. It is a measure of how quickly the average cited article, in a particular journal is cited. In contrast, the *cited half-life* is the median age of the articles cited in the current year.

Among the traditional journal citation measures, the well-known *impact factor* has been

regarded as the best instrument for the evaluation of the quality of scientific journals. It is usually used by researchers deciding where to publish and what to read, by editors and publishers as a means to evaluate and promote their journals, and by tenure and promotion committees laboring under the assumption that publication in a higher impact factor journal represents better work [149, 235, 473]. This widespread usage is obviously due to its simplicity. Despite its advantages, the *impact factor* has not been spared from criticism [10, 47, 127, 144, 146, 184, 331, 407, 457, 449, 485]. Main points of consideration regarding methodological aspects in the calculation of this index include the lack of assessment of the quality of citations, the inclusion of self-citations, and the poor comparability between different scientific fields.

4.5.2 Bibliometric measures that assess the quality of citations

The metric used to quantify the importance of scientific publications has been largely based on integer counting of their citations. Although, the number of citations gives a direct approximation of a journal's importance, some situations reflect that citations do not seem to provide a full picture of the prestige of a journal. In this context, new measures, which concern more about who actually cited the journal and the prestige they have transferred to the cited journal, are proposed. The common point in most of these new measures is the assessment of the quality of citations received by a journal [363]. The quality of citations can be estimated analyzing the networks of scientific papers with sophisticated mathematical algorithms. The PageRank algorithm [63] has been proposed as an appropriate model for the evaluation of the quality of citations in scientific journals. Some PageRank-inspired indicators have been introduced in the literature [42, 47, 88, 186, 191, 368]. In the following, the Eigenfactor metrics [42] and the Scimago Journal Rank indicators [186, 191] are detailed.

Eigenfactor score This measure [42] represents the journal's total importance to the scientific community. The *Eigenfactor score* weights journal citations by the influence of the citing journals. As a result, a journal is influential if it is cited by other influential journals. It ranks a journal according to the sum of normalized citations received from other journals weighted by the influence of the citing journals. It uses a target window of five years which allows, in general, a broader evaluation of journal citations, in particular for disciplines with longer cited lives. The *Eigenfactor score* is a size-dependent measure: with all else equal, bigger journals will have larger *Eigenfactor scores*, since they have more articles and hence we expect them to be cited more often. Finally, the *Eigenfactor scores* are scaled so that the sum of the scores of all journals is 100. This approach is thought to be more robust than the *impact factor*, which purely counts the number of citations without considering the significance of those citations.

Article Influence score It measures the average influence, per article, of the papers in a journal. It is calculated as the *Eigenfactor score* divided by the number of articles published by the journal over the five-year target period. The *Article Influence scores* are normalized and can be comparable to the *impact factor*.

SJR indicator The *Scimago Journal Rank* [186] is a measure of scientific influence of scholarly journals that accounts for both the number of citations received by a journal and the importance or prestige of the journals where such citations come from. It is a size-independent indicator and its values order journals by their average prestige per article and can be used for journal comparisons in science evaluation processes. The *SJR indicator* assigns different values to citations depending on the importance of the journals where they come from, that is, citations coming from highly important journals will be more valuable and hence will provide more prestige to the journals receiving them. The calculation of the *SJR indicator* is very similar to the *Eigenfactor score*, with the former being based on the Scopus database and the latter on the Web of Science database.

SJR2 indicator This indicator [191] takes into account not only the prestige of the citing scientific journal but also its closeness to the cited journal using the cosine of the angle between the vectors of the two journals' cocitation profiles. The *SJR2 indicator* was designed to weight the citations according to the prestige of the citing journal, also taking into account the thematic closeness of the citing and the cited journals. The procedure does not depend on any arbitrary classification of scientific journals, but uses an objective informetric method based on cocitation. It also avoids the dependency on the size of the set of journals, and endows the score with a meaning that other indicators of prestige do not have.

4.5.3 Bibliometric measures that correct for differences among fields

It is well known that in some scientific fields the average number of citations per publication is much higher than in other scientific fields. This is due to differences among fields in the average number of cited references per publication, the average age of cited references, and the degree to which references from other fields are cited, among others. The first suggestions on how to control for these factors and calculate normalized citation rates were made in the 1980s [404, 463]. In general, all proposed normalization methods are computed by dividing the actual number of received citations for a group of publications with the number of citations that could be expected for similar publications.

Two well-known field normalized citation scores, called the *crown indicator* [330] and the *mean normalized citation score* [469], were developed to correct for the citation differences among fields. Similarly, the *normalized mean citation rate* [182, 402] was also proposed for the above purpose. Other works tried to construct a field-independent indicator dividing the *h-index* by the average number of citations per paper [227], dividing the actual number of citations of each paper of a researcher by the average number of publications in the field [373], dividing the journal's citation count per paper by the citation potential in its subject field [329], and assigning weights to journal citations according to the journal's average number of references per article [492].

crown indicator The *crown indicator* [330] relies on a normalization mechanism that aims to correct for the citation differences among fields. The normalization mechanism basically

works as follows. Given a set of publications, the number of citations that each publication has received, is counted. Also, the expected number of citations of each publication is determined. The expected number of citations of a publication equals the average number of citations of all publications of the same document type (i.e., article, letter, or review) published in the same field and in the same year. In this context, the *crown indicator* is calculated by dividing the average number of received citations for a group of publications with the average number that could be expected for publications of the same type, from the same year, published in journals within the same field. Mathematically, it is defined as

$$crown\ indicator = \frac{CPP}{FCS_m} \tag{4.18}$$

where CPP is the average number of citations per publication, without self-citations, and FCS_m is the mean field-normalized citation score which is calculated using the same publication and citation counting procedure as in the case of CPP. The normalization mechanism of the crown indicator has been criticized in [300, 350]. These authors have argued in favor of an alternative mechanism which calculates for each publication the ratio of its actual number of citations and its expected number of citations and then takes the average of the ratios that one has obtained.

mean normalized citation score The *crown indicator* was recently modified into the *mean normalized citation score* [469] in order to overcome some drawbacks of the *crow indicator*. The most important drawback is that citation rates are not normalized on the level of individual publications. This way of calculating gives more weight to older publications (particularly reviews), published in fields with dense citation traffic. In order to give each publication equal weight the normalization should take place on the level of the individual publication. Instead of first calculating the actual average citation rate, and then divide that with the average expected citation rate, each publication is normalized individually. It is mathematically defined as

$$mean\ normalized\ citation\ score = \frac{1}{n}\sum_{i=1}^{n}\frac{cit_i}{exp_i} \tag{4.19}$$

where n is the number of papers, cit_i is the number of citations of paper i, and exp_i is the expected number of citations of paper i. The *mean normalized citation score* has the advantage of being mathematically consistent while the previous *crown indicator* was not. Comparing the *crown indicator* and the *mean normalized citation score*, it can be seen that the *crown indicator* normalizes by calculating a ratio of averages while the *mean normalized citation score* normalizes by calculating an average of ratios. Hence, while the *crown indicator* performs a normalization at the level of an oeuvre as a whole, the *mean normalized citation score* performs a normalization at the level of the individual publications in an oeuvre.

4.6 Bibliographic databases

Bibliographic databases, also called citation databases, are used to combine information related to bibliographic productivity and to facilitate the identification of authors of publications and sources of publication citations. Historically, the most common applications have been calculation of the scientific relevance of scholarly journals and evaluation of researcher productivity. A large number of thematic citation databases are available, but their coverage is limited to specific scientific areas. Other databases are more general and have been constructed to cover the overall academic productivity. *Web of Science* [1], *Scopus* [2] and *Google Scholar* [3] are the most well-known databases.

Garfield founded the Institute for Scientific Information (ISI) and constructed the first citation databases for combining information on publications and associated citations for a defined set of scientific journals [170] in 1955. ISI databases have been the most generally accepted data sources for bibliometric analysis. They have built a reputation as the oldest citation resources, containing the most prestigious academic journals [347]. Nowadays, Thomson Reuters, one of the world's largest information companies, continues the work initiated by the Institute for Scientific Information, developing the new *Web of Science* platform.

Two alternatives to *Web of Science* that are growing in popularity are *Scopus*, provided by Elsevier, and *Google Scholar*, provided by Google Inc. After their remarkable introduction in 2004, the challenge for *Scopus* and *Google Scholar* are to position themselves as citation resources in a market where *Web of Science* held the monopoly [198, 307]. Since *Scopus* and *Google Scholar* began competing with *Web of Science*, a large number of comparisons of coverage and bibliometric measures calculated on the basis of the three citation databases have been published [26, 31, 32, 169, 198, 228, 230, 231, 232, 321]. Although a comprehensive review is beyond the scope of this chapter, a few comparison examples are detailed below.

Recent works have compared citation resources across different parameters. On the one hand, Adriaanse and Rensleigh [11] compared citation counts, multiple copies and inconsistencies encountered across the three citation resources *Web of Science*, *Scopus* and *Google Scholar*. Data from the South African scholarly environmental sciences journals for the year range 2004-2008 were extracted from the three citation resources and compared. The total citation counts indicated that *Web of Science* retrieved the most citation results, followed by *Google Scholar* and then *Scopus*. *Web of Science* performed the best with total coverage of the journal sample population and also retrieved the most unique items. The investigation into multiple copies indicated that *Web of Science* and *Scopus* retrieved no duplicates, while *Google Scholar* retrieved multiple copies. *Scopus* delivered the least inconsistencies regarding content verification and content quality compared to the other two citation resources. Additionally, *Google Scholar* also retrieved the most inconsistencies, with *Web of Science* retrieving more inconsistencies than *Scopus*. Examples of these inconsistencies include author spelling and sequence, volume and issue number. On the other hand, Bartol *et al.* [35] analyzed all documents and citations received by authors who were actively engaged in research in Slovenia between 1996 and 2011. They showed that *Scopus* leads over *Web of*

Science in indexed documents as well as citations in all research fields. This is especially evident in social sciences, humanities, and engineering and technology. The least citations per document were received in humanities and most citations in medical and natural sciences, which exhibit similar counts. Finally, Chirici [91] compared the scientific productivity of the Italian forestry community for 1996-2010 using some bibliometric measures. Results showed that mean number of publications, mean number of citations and *h-index* values calculated by *Web of Science* and *Scopus* were not statistically different. He also found that *Web of Science* has a more complete and wider coverage for the analyzed authors than *Scopus*.

It needs to be underlined that the coverage, scope, search functionalities and analysis tools in all databases are constantly evolving; therefore, the information in the literature can only refer to the coverage that exists at the time of the analysis. A clear winner among citation resources is not possible to reveal since the relative advantages of one over the others depend on what it is specifically wanted to analyze. Some authors [147, 193, 279, 284, 288, 438] noted that the coverage in *Scopus* outperformed *Web of Science* in specific disciples. Other works [31, 153, 270, 320, 392] showed that *Google Scholar* computes significantly higher indicators' scores than *Web of Science* and *Scopus*. In contrast, *Google Scholar*'s lack of quality control limits its use as a bibliometric tool because of non-scholarly sources, erroneous citation data and errors of omission and commission when using search features [323]. The search results from *Google Scholar* are very noisy and therefore require considerable difficult and time-consuming filtering to obtain usable information, especially for evaluation purposes. Also, *Google Scholar* is the only citation resource retrieving multiple copies (duplicates and triplicates). The poor capability of consolidating matching records causes *Google Scholar* to inflate the citation hits which does not give a true reflection of the citation count [229].

Although the three sources have different goals and contents, they all track citations that are potentially useful for bibliometric studies [323]. Despite the differences, all databases are highly correlated and comparable in terms of rankings at the macro level [31, 33]. A considerable overlap among *Web of Science*, *Scopus* and *Google Scholar* in terms of content is found by different works [11, 408].

4.6.1 Web of Science

Web of Science is the most comprehensive and versatile research platform available. It offers researchers, administrators, faculty, and students with quick, powerful access to the world's leading citation databases and gives them powerful tools to search, track and measure research publications. The careful selection process ensures users to get the most reliable, integrated, multidisciplinary information from the global research community. Its content and tools are trusted by more than 7,000 of the world's leading scholarly institutions responsible for scientific policy making.

Underneath the new *Web of Science* umbrella, a user-friendly discovery environment provides access to many resources like the *Web of Science core collection*, including the well-known *Science Citation Index Expanded*, *Social Sciences Citation Index*, and *Arts and Humanities Citation Index*, and other resources like *Conference Proceedings Citation Index*,

Book Citation Index, *Index Chemicus* and *Current Chemical Reactions*. The *Web of Science core collection* contains more than 55 million records from the top journals, conference proceedings, and books in the sciences, social sciences, and arts and humanities. It provides an authoritative and multidisciplinary coverage, dating back to 1900, from more than 13,000 high impact research journals worldwide, 160,000 conference proceedings and 60,000 books associated with more than 250 disciplines.

A short description of the main *Web of Science core collection* resources are as follows:

- *Science Citation Index Expanded* focus on bibliographic and citation information from over 8,500 of the world's leading scientific and technical journals across 150 disciplines. It has more than 43 million records and delivers comprehensive backfile and cited reference data from 1900 to the present.

- *Social Sciences Citation Index* contains essential data from 3,000 of the best social sciences journals across 50 disciplines. It stores more than 7 million records and its range of coverage is from the year 1956 to the present day.

- *Arts and Humanities Citation Index* indexes over 1,700 arts and humanities journals. It contains more than 4 million records and provides backfiles to 1975.

- *Conference Proceedings Citation Index* helps researchers access the published literature from the most significant conferences worldwide. It covers more than 160,000 conference proceedings starting from 1990 to the present day. It holds more than 8 million records.

- *Book Citation Index* indexes over 60,000 editorially selected books with 10,000 new books added each year. It introduces more than 15 million of new cited references from 2005 to the present.

- *Index Chemicus* provides access to the chemical compound information, covering more than 100 of the world's leading organic chemistry journals. It houses more than two million compounds from 1993 to the present.

- *Current Chemical Reactions* contains over one million reactions. Each reaction provides complete bibliographic data which is available back to 1840.

Beyond the *Web of Science core collection*, the following additional databases support the *Web of Science*. The *Chinese Science Citation Database* covers over 1,200 top scholarly publications from China. It contains a multidisciplinary coverage to 1989 of many disciplines, including nearly 2 million article records and more than 13 million citations. The *SciELO Citation Index* discovers new insights from research in Latin America, Spain, Portugal, the Caribbean and South Africa. It provides access nearly 700 titles and stores over 4 million cited references. The *KCI Korean Journal Database* lets researchers discover new insights from research emanating from South Korea. It incorporates nearly 1,500 scholarly journals to the *Web of Science*. Also, *Derwent Innovations Index* facilitates rapid, precise patent and citation

searches of inventions in chemical, electrical, electronic, and mechanical engineering. It covers over 16 million records from 41 worldwide patent-issuing authorities and provides backfiles to 1963. Finally, *Current Contents Connect* provides easy Web access to complete tables of contents, abstracts, bibliographic information, and abstracts from about 8,000 journals and 2,000 books. It contains over 20 million records and includes pre-published electronic journal articles and links to the full text.

Other discipline-based databases are also incorporated into the *Web of Science*. *BIOSIS Previews* uncovers relevant coverage in the life sciences research. It stores data over 5,200 journals, as well as academic books, abstracts, published theses, conference proceedings, bulletins, monographs, and technical reports and contains over 21 million records to 1926. *Zoological Record* is the world's oldest continuing database of animal biology. It has over 4 million records to 1864. *CAB Abstracts* is the most comprehensive source of international research information in agriculture. It indexes over 7,500 journals, as well as non-journal literature, contains more than 7 million records and provides backfiles to 1910. *Food Science and Technology Abstracts database* provides thorough coverage of pure and applied research in food science, food technology, and food-related human nutrition. It has over one million records from more than 4,600 serial publications and patents from around the globe. *Inspec* is a comprehensive index to literature in physics, electrical/electronic technology, computing, control engineering, information technology, manufacturing, production and mechanical engineering. It covers over 13 million bibliographic records from publications worldwide which is available back to 1898. Finally, *Medline* is the premier bibliographic database of the U.S. National Library of Medicine, covering biomedicine and the life sciences, bioengineering, public health, clinical care, and plant and animal science. It houses over 17 million records from publications worldwide, including over 5,300 journals in 30 languages, plus a select number of relevant items from newspapers, magazines, and newsletters.

Some advantages of using the *Web of Science* platform are detailed as follows:

- Many qualitative and quantitative factors are taken into account when evaluating journals, conferences and books for coverage in *Web of Science*. They are constantly under review, ensuring that they are maintaining high quality standards. Also, the journals indexed in *Web of Science* are the same journals found in *Journal Citation Reports (JCR)* and it is probably considered the premiere source by many to consult when looking for journal impact.

- It tracks over a century of vital data and provides high quality data to perform comprehensive analysis. More backfiles give the power to conduct deeper searches and track trends through time.

- It continues to expand its coverage by adding new regional databases, including specific research in China, South Korea, Latin America, Spain, Portugal, the Caribbean and South Africa, and subject-specific databases, like *BIOSIS Previews*, *Inspec*, and *Medline*.

- It indexes bibliographic data from cover-to-cover so it is possible to access every significant item from journals, conferences and books, including original research articles, reviews, editorials, chronologies, abstracts, proceeding papers, book chapters and more. Also, it provides access to publishers' full-text documents.

- It has complex and focused search options. Users can navigate forward and backward through the literature, and search all disciplines and time periods. It enhances the power of cited reference searching by searching across disciplines for all the related articles that have cited references in common.

- Author identification tools are available. It locates documents written by the same authors in a simple, single search, eliminating the problems of similar author names or several authors with the same name. It also supports Unicode expanding search capabilities to Chinese characters, among others.

- It has many insightful analysis options to discover hidden trends and patterns, gain insight into emerging fields of research, and identify leading researchers, institutions, and journals. Visual and graphical representations of citation activity are also available.

- It includes *Essential Science Indicators* which can determine the influential individuals, institutions, papers, publications, and countries in a field of study. This unique and comprehensive compilation of science performance statistics is an ideal analytical resource for policymakers.

- It allows the integration with EndNote, providing free access to all academic users in order to store, organize and share their references online. ResearcherID and RSS feeds are also integrated in the *Web of Science* platform.

In summary, *Web of Science* collects bibliographic and citation information which has met the high standards of an objective evaluation process, eliminating clutter and delivering accurate, meaningful and timely data. It contains more than 90 million records, more than one billion cited references, and indexes around 65 million items per year. It also provides comprehensive backfile and cited reference data from 1900 to the present. The multidisciplinary coverage of the *Web of Science* encompasses over 20,000 journals within 3,300 publishers, 160,000 conference proceedings including over 8 million conference papers, and 60,000 scholarly books. The coverage includes: the sciences, social sciences, arts, and humanities, and goes across more than 250 disciplines.

4.6.2 Scopus

Scopus is a bibliographic database containing abstracts and citations for multidisciplinary literature. It is designed to serve the research information needs of researchers, educators, administrators, students and librarians across the entire academic community. Like *Web of Science*, *Scopus* features smart tools to track, analyze and visualize research. Whether

searching for specific information or browsing topics, authors or journals, *Scopus* provides precise entry points to peer-reviewed literature in the fields of life sciences, physical sciences, health sciences and social sciences and humanities. These fields can be further divided into 27 major subject areas and more than 300 minor subject areas.

Updated daily, *Scopus* includes bibliographic information of over 22,000 peer-reviewed journals from more than 5,000 international publishers. In addition, it includes over 6.5 million conference papers from over 18,000 worldwide events, more than 50,000 books, 545 million scientific web results and 25.2 million patents records from the most important five patent offices. Depth of *Scopus* coverage is not as impressive as the width because many journals and conferences are only covered for the last few years.

Although records in the database go as far back as 1823, references do not appear in those records until 1996. In this context, *Scopus* collects 54 million records, including 33 million records with references back to 1996 and 21 million records pre-1996 with no references. Bibliometric calculations based on them are only available from publications since 1996, resulting in very skewed bibliometric measures for researchers with longer careers than this.

Some advantages of using the *Scopus* platform are as follows:

- It provides comprehensive journal and country ranking data on its *Scimago Journal and Country Rank*. It includes scientific indicators, which can be used to assess and analyze scientific domains, developed from the information contained in the *Scopus* database.
- An independent and international *Content Selection and Advisory Board* is established to prevent a potential conflict of interest in the choice of journals to be included in the database and to maintain an open and transparent content coverage policy.
- It identifies and matches organizations and collaborators with their research output using identifier tools. It clarifies their identity through integration with ORCID.
- Search and filter procedures are very fast, allowing searches of scientific web pages, author homepages and university sites through *Scirus*.
- The retrieve results can be download in different formats using the *Quosa Document Download Manager*. Also, it is possible to export data to reference managers such as Mendeley, RefWorks and EndNote.

4.6.3 Google Scholar

Google Scholar is a web search engine providing free access to the world's scholarly literature. It primarily searches academic papers from most major academic publishers and repositories worldwide, institutional and individual bibliographic databases. Although it does not specifically list the sources of its data, it is designed to be as comprehensive as possible.

The size of *Google Scholar*'s database is not published. Despite this, researchers estimated it to contain roughly 160 million documents. It includes journal and conference papers, theses,

dissertations, academic books, pre-prints, abstracts, technical reports and other scholarly literature. It also includes court opinions and patents. Shorter articles, such as book reviews, news sections, editorials, announcements and letters, may or may not be included. Untitled documents and documents without authors are usually not included. Website URLs that aren't available to *Google* search robots are, obviously, not included either.

Although *Google Scholar* citations are usually considered of comparable quality and utility to commercial databases, they have been found to be sometimes inadequate since they are vulnerable to spam. Results often contain duplicates of the same article due to the wide range of sources. Also, it has been found to include duplicate citations, part because of inclusion of multiple copies. In this context, citation counts from *Google Scholar* should only be used with care especially when used to calculate performance metrics.

The most important advantage of using *Google Scholar* is that it stands out in its coverage of conference proceedings as well as international, non-English language journals, and its greater coverage includes some items that are not found in the other databases. It also covers not only journals but academic grey literature and electronic-only publications.

Finally, Table 4.2 shows a short comparison among these three bibliographic databases. Although *Google Scholar* does not provide information associated with many aspects, it has the highest number of records and it is the unique database which provides free access to researchers. *Web of Science* is the oldest resource and excels in the citation coverage. In contrast, *Scopus* overcomes the number of covered journals and patents.

Table 4.2: Short comparison among three bibliographic databases.

Features	*Web of Science*	*Scopus*	*Google Scholar*
Institution	Thomson-Reuters (USA)	Elsevier (Holland)	Google (USA)
Date	1960s	2004	2004
Records	90 million	54 million	160 million
Journals	20,000	22,000	Unknown
Publishers	3,300	5,000	Unknown
Conferences	8 million	6.5 million	Unknown
Books	60,000	50,000	Unknown
Patents	14.3 million	25.2 million	Unknown
References	from 1900	from 1996	Unknown
Disciplines	> 250	> 300	Unknown
Licence	Unfree	Unfree	Free
Update	Weekly	Daily	Daily

Part III

Data Mining in Research Evaluation

Chapter 5

Predicting citation counts using supervised algorithms

5.1 Introduction

Publishers nowadays face the problem of deciding which of the many papers they receive are of higher quality for publication in their journals. The current method used for article assessment is peer review. Although if used properly peer review is assumed to be the most reliable system, it is slow, expensive and unwieldy [93, 338, 393]. Other authors contest this appraisal [195, 212]. This difference of opinion among authors has led to the development of several quantitative metrics associated with scientific production. One such metric is citation count. Citation count is the number of citations received by a paper in a period of time. Although citations are a measure of visibility, they can be considered as an indirect measure of article quality. The aim of this measure is to mirror the impact and quality of papers [52].

This chapter is focused on the construction of predictive models to forecast the citation count of papers published in the *Bioinformatics* journal within four years after publication. In this context, several researchers have investigated the prediction of citation count. Their work differs primarily as regards the prediction time horizon for the citation count and the predictive features used.

Some papers predict the number of citations using information gathered after publication. On the one hand, Brody *et al.* [64] used download data as predictive features. Authors found a correlation of ρ=0.44 between number of citations and total downloads over the two years after publication. In this context, Perneger [361] found a correlation of ρ=0.54 between the total number of citations after five years and the number of views of the full-length HTML version of the article during the first week after publication. Also, Watson [472] showed that the number of downloads per day over the first 1,000 days after publication was correlated well with the number of citations per year. On the other hand, Castillo *et al.* [80] used the number of citations, the authors' reputation and the source of the paper citations as predictive features. Lokker *et al.* [296] used features related to the article and journal,

like number of authors, pages, references and so on. The above three works used measures taken after the paper was published to predict its citation count in the future. The main disadvantage of using these features is that the required values are not available until several months after publication. In contrast, others papers attempt to forecast citation count with the information available at the time of publication. Fu and Aliferis [164] predict citation count within ten years after publication with bibliometric information (number of articles for the first author, number of citations for the first author, number of authors, number of institutions and so on), the journal impact factor and the content of the article (title, abstract and MeSH terms). All these features are available at the time of publication. Support vector machine classification models were used as the learning algorithm. Predictions were made for a simple binary response variable that is defined by a set of citation thresholds to determine if an article is labeled positively or negatively. For a given threshold t, a positive label means that an article received at least t citations within ten years after publication. These thresholds were 20 (mildly influential), 50 (relatively influential), 100 (influential) and 500 (extremely influential). Depending on the threshold used, the models output area under the ROC curve (AUC) values ranging from 0.857 to 0.918.

As in Fu and Aliferis [164], this chapter also deals with the response variable as a discrete variable. Unlike them, the variable that counts the number of citations is discrete rather binary but, taking three possible values (*few*, *some* and *many* citations). This leads to the use of classification methods rather than regression models to predict citation counts. Unlike Fu and Aliferis [164] that use only support vector machines, several classification methods are taken into account and it is analyzed which one provides better predictions for the problem. Moreover, the proposed models will be constructed especially to predict annual time horizons (each of the first four years after publication) and for each *Bioinformatics* journal section. The information required from each article is its abstract content and the number of two-week periods after publication. Hence, as opposed to other previous models described above that require information that is not available until after publication, these predictions will be available at publication time. Also, the information output by the model will be exploited, like the identification of key features (e.g. words in the abstract) that increase the chances of citation. This method can actually inform publishers about which articles will have a bigger impact in the future before they are published. This work appears in the published paper [223].

Chapter outline

Section 5.2 presents the prediction of future citation count of Bioinformatics papers within four years of publication. It includes the dataset compilation, the data distribution and the predictive models which are learned from data. It also shows a prediction example using the models which have the best performance. Finally, Section 5.3 discusses the main results and conclusions achieved in this chapter.

5.2 Predicting citation count of Bioinformatics papers

Two different types of predictive models are learned in this chapter. Global models attempt to predict the number of citations received by an article within each of the four years after publication, using information on all papers published in *Bioinformatics* over three years, from January 1, 2005 to December 31, 2007. Specific models have the same objective but, in this case, they use the information related to articles published within a specific *Bioinformatics* journal section.

Different phases are required for the dataset construction. The collection of abstracts published in *Bioinformatics* is the starting point for the construction of predictive models.

5.2.1 Dataset compilation

Bioinformatics is selected as the journal for this study. The basic elements of this work are the abstracts published in the *Bioinformatics* journal sections (*Data and Text Mining, Databases and Ontologies, Gene Expression, Genetics and Population Analysis, Genome Analysis, Phylogenetics, Sequence Analysis, Structural Bioinformatics* and *Systems Biology*) from 2005 to 2007. Before that date, no such sections existed. These abstract are collected from the journal website (http://bioinformatics.oxfordjournals.org/). Once this information is downloaded, the abstracts, the journal section and the number of two-week periods from the beginning of the year to the publication date are stored in a database designed for this purpose.

The next step is to extract tokens from the abstracts. To do this, a list of tokens ordered by frequency of occurrence in the abstract set is achieved. This list is composed of one-, two- and three-word tokens. Then, the list is filtered to reduce the large number of different tokens. The proposed filter is based on removing tokens that appear only occasionally in the abstract set. In this way, tokens that have a frequency of occurrence of less than three will be removed. After that, some tokens that are repeated frequently and are irrelevant to the case study are eliminated. For example, prepositions and articles are classic examples of stopwords. Generally, these tokens appear in all abstracts, and play no role in building the predictive model. Finally, each token is associated with their morphological root using Lucene software. The last step of the dataset compilation was to download the number of citations received by each article within each year after publication until December 31, 2008 from the Web of Science.

The dataset structure is different depending on the model to be built. The data set structure of global models is made up of the *Section*, *Date*, *Token-1*, ..., *Token-n* features and *Citation* variable; whereas the specific models have the same structure except for the *Section* feature, which is constant.

- *Section* can take the values: *1-Data and Text Mining, 2-Databases and Ontologies, 3-Gene Expression, 4-Genetics and Population Analysis, 5-Genome Analysis, 6-Phylogenetics, 7-Sequence Analysis, 8-Structural Bioinformatics* and *9-Systems Biology.* These values correspond to the different *Bioinformatics* journal sections.

- The feature *Date* refers to the number of two-week periods from the beginning of the year to the publication date. It can take the values {1,2,...,24}.

- *Token-i* are the features that belong to the final list of the tokens extracted from abstracts. These features are binary, and take the value 1 or 0 depending on whether or not the token is present in the selected abstract.

- *Citation* variable corresponds with the class label. It can take the values {*few*, *some*, *many*}. The first value, *few*, describes papers that receive at most one citation in a specific year according to the Web of Science. The value *some* applies to papers that receive 2, 3 or 4 citations in a year. And finally, the value *many* refers to papers that receive a number of citations equal to or greater than five.

5.2.2 Data distribution

The number of articles selected to build the predictive models varies depending on the prediction year. To construct the models assigned to the *first-* and *second-year*, articles published in the years 2005, 2006 and 2007 were used. On the other hand, the models for the *third-year* used papers published in 2005 or 2006, and finally, the predictive models for the *fourth-year* used articles published in 2005 only. Clearly, the longer the prediction horizon is the fewer papers are used to induce the models.

Table 5.1 shows the distribution of the articles selected in this research. It illustrates the number of papers belonging to a journal section in a particular year. Furthermore, it shows the distribution associated with each value of the class to be predicted. The value (*All; Second-year; Total*) shows that 1086 articles are available to induce the global models in the *second-year*. According to the number of citations received, these articles are further divided into *few* (388), *some* (432) and *many* (226). Table 5.1 also lists the number of papers

Table 5.1: Distribution of the data (papers), according to nine journal sections and citation count (*few*, *some* and *many*) across the four-year time horizon.

	First-year				*Second-year*				*Third-year*				*Fourth-year*			
	Total	*f*	*s*	*m*	Total	*f*	*s*	*m*	Total	*f*	*s*	*m*	Total	*f*	*s*	*m*
S1	**88**	81	7	0	**88**	39	31	18	**50**	10	17	23	**24**	8	7	9
S2	**37**	32	3	2	**37**	13	14	10	**25**	6	8	11	**15**	6	3	6
S3	**283**	253	26	4	**283**	107	114	62	**192**	38	66	88	**107**	23	37	47
S4	**46**	41	5	0	**46**	19	16	11	**27**	7	9	11	**21**	1	11	9
S5	**103**	93	9	1	**103**	26	46	31	**82**	19	27	36	**53**	11	22	20
S6	**28**	23	4	1	**28**	6	14	8	**20**	2	9	9	**11**	2	4	5
S7	**190**	170	16	4	**190**	73	77	40	**141**	49	46	46	**82**	22	31	29
S8	**150**	130	19	1	**150**	49	55	46	**103**	22	35	46	**54**	6	13	35
S9	**161**	140	20	1	**161**	56	65	40	**100**	14	32	54	**53**	3	12	38
All	**1086**	963	109	14	**1086**	388	432	266	**740**	167	249	324	**420**	82	140	198

S1 (Data and Text Mining), S2 (Databases and Ontologies), S3 (Gene Expression), S4 (Genetics and Population Analysis), S5 (Genome Analysis), S6 (Phylogenetics), S7 (Sequence Analysis), S8 (Structural Bioinformatics), S9 (Systems Biology)

used in the specific models. For example, the models associated with *5-Genome Analysis* and *third-year* use 82 papers. Furthermore, section *3-Gene Expression* accounts for 26.01% of articles used in the *first-year* prediction models, whereas the sections with fewer papers in the *first-year* are *2-Databases and Ontologies* (3.41%), *4-Population Genetics and Analysis* (4.23%) and *6-Phylogenetics* (2.58%). The sections with more and fewer papers are the same across all years.

5.2.3 Predictive models

Different classifiers (naive Bayes, K2 algorithm, logistic regression, C4.5 algorithm and k-nearest neighbor algorithm) have been used to learn the proposed global and specific models. Before running the above methods, feature selection is performed. To determine whether all features are equally important or necessary to discriminate between the values {*few*, *some*, *many*}, correlation-based feature selection is run. The objective of feature selection is to build parsimonious models. Features that are irrelevant or redundant will not appear in these models. Also, k-fold cross-validation is used as the procedure for estimating the probability of models classifying new cases according to the value of the predictive features.

Several global models have been constructed for predicting the citation count of all the articles within four years after publication. Each model is associated with one of the four years to be predicted and one of the five supervised classification methods studied. Table 5.2 shows the results for each model. These results could be better since apart from *first-year* models, model accuracy is less than 80%. There are some classification methods that provide better results than others. In this case, Bayesian classifiers have a higher average success rate within the four years (naive Bayes-73.40% and K2-70.37%), whereas logistic regression (65.85%), decision trees (60.15%) and K-NN (56.47%) yield the worst results. Although the *first-year* model has a much higher success rate than the models for the other years, the results are not satisfactory. This is because most cases belong to the *few* class (Table 5.1), and this is an obstacle to learning about the *some* and *many* classes since models avoid classifying cases into these classes. The C4.5 and K-NN methods especially tend to make this error for the *first-year* time horizon, whereas Bayesian classifiers and logistic regression are not prone to this error.

Table 5.2: Accuracy and standard deviation of global models.

	All journal sections			
	First-year	*Second-year*	*Third-year*	*Fourth-year*
NB	91.4 ± 1.62	57.4 ± 6.08	68.9 ± 5.07	75.9 ± 6.39
K2	89.7 ± 2.54	57.4 ± 5.38	65.3 ± 4.48	69.1 ± 6.83
LR	84.7 ± 3.95	56.6 ± 2.75	59.3 ± 5.56	62.8 ± 7.20
C4.5	88.2 ± 0.47	48.8 ± 4.02	48.6 ± 4.69	55.0 ± 4.74
K-NN	88.5 ± 0.73	44.6 ± 4.72	38.5 ± 4.55	54.3 ± 4.95

NB (naive Bayes), K2 (K2 algorithm), LR (logistic regression),
C4.5 (C4.5 algorithm), K-NN (k-nearest neighbor algorithm)

Table 5.3: Accuracy and standard deviation of specific models. Numbers in boldface represent an average success rate better than 95%.

Methods	*S1*	*S2*	*S3*	*S4*	*S5*	*S6*	*S7*	*S8*	*S9*
1st-year	7 features	7 features	30 features	5 features	11 features	18 features	33 features	54 features	48 features
NB	**96.6 ± 5.42**	91.9 ± 12.1	94.0 ± 4.42	**95.6 ± 8.45**	93.2 ± 4.73	**100 ± 0.00**	**95.8 ± 3.58**	**97.3 ± 3.42**	**96.9 ± 3.32**
K2	**95.6 ± 5.74**	92.5 ± 12.1	94.0 ± 3.26	**98.0 ± 6.32**	92.3 ± 4.09	**96.7 ± 10.5**	**96.3 ± 3.55**	**95.3 ± 4.50**	**96.3 ± 3.20**
LR	**98.9 ± 3.53**	86.5 ± 17.7	90.8 ± 2.55	**95.7 ± 8.46**	94.2 ± 5.06	92.9 ± 14.0	94.7 ± 4.36	92.0 ± 4.26	91.9 ± 4.81
C4.5	92.0 ± 5.42	86.5 ± 13.2	90.1 ± 2.22	89.1 ± 10.5	90.3 ± 0.48	82.1 ± 17.6	89.5 ± 2.53	86.7 ± 3.19	88.2 ± 3.59
K-NN	92.0 ± 5.42	86.5 ± 13.2	89.4 ± 0.22	89.1 ± 10.5	90.3 ± 0.48	82.1 ± 17.6	89.5 ± 0.00	86.7 ± 0.00	87.0 ± 1.65
2nd-year	71 features	50 features	187 features	52 features	109 features	29 features	128 features	134 features	106 features
NB	82.9 ± 12.3	89.2 ± 21.9	86.2 ± 6.23	84.8 ± 14.1	86.4 ± 6.59	89.3 ± 16.1	85.3 ± 6.65	88.7 ± 7.20	87.0 ± 8.52
K2	69.3 ± 15.4	72.5 ± 23.3	75.3 ± 7.39	78.0 ± 15.3	70.6 ± 14.5	70.0 ± 24.6	80.0 ± 7.94	68.0 ± 12.1	80.8 ± 11.5
LR	**96.6 ± 5.43**	94.6 ± 12.4	90.1 ± 5.42	**95.7 ± 9.62**	**95.1 ± 6.93**	92.9 ± 14.0	**95.3 ± 5.87**	94.7 ± 7.60	93.2 ± 6.17
C4.5	56.8 ± 11.4	37.8 ± 11.3	53.7 ± 6.06	58.7 ± 22.8	61.2 ± 14.9	57.1 ± 30.6	58.9 ± 8.03	54.0 ± 6.90	56.5 ± 9.60
K-NN	53.4 ± 9.34	73.0 ± 20.1	62.2 ± 8.94	50.0 ± 17.1	63.1 ± 15.5	60.7 ± 24.6	57.4 ± 7.91	72.7 ± 15.5	60.2 ± 9.35
3rd-year	58 features	37 features	143 features	37 features	87 features	18 features	149 features	109 features	83 features
NB	90.0 ± 14.1	80.0 ± 21.9	85.9 ± 6.81	88.9 ± 22.5	**95.1 ± 6.34**	85.0 ± 24.1	86.5 ± 7.14	89.3 ± 11.0	84.0 ± 11.7
K2	72.0 ± 21.5	73.3 ± 25.1	70.8 ± 7.23	63.3 ± 28.1	72.9 ± 14.4	90.0 ± 21.1	71.7 ± 14.2	77.0 ± 15.4	77.0 ± 14.9
LR	94.0 ± 9.78	**100 ± 0.00**	90.6 ± 7.74	92.6 ± 14.0	93.9 ± 6.66	**100 ± 0.00**	**95.7 ± 5.09**	93.2 ± 6.70	93.0 ± 8.20
C4.5	50.0 ± 14.1	48.0 ± 12.3	66.7 ± 7.86	48.1 ± 21.4	54.9 ± 9.78	70.0 ± 35.0	58.2 ± 9.23	58.3 ± 11.1	57.0 ± 11.6
K-NN	64.0 ± 15.8	52.0 ± 12.3	54.7 ± 8.70	92.6 ± 21.1	58.5 ± 14.2	75.0 ± 26.3	58.9 ± 9.12	65.1 ± 11.4	60.0 ± 6.70
4th-year	32 features	12 features	91 features	25 features	58 features	10 features	80 features	39 features	22 features
NB	91.7 ± 21.1	80.0 ± 24.1	86.0 ± 4.53	90.5 ± 18.0	90.6 ± 9.93	81.8 ± 42.2	93.9 ± 10.6	87.0 ± 12.0	90.6 ± 13.9
K2	73.3 ± 33.5	70.0 ± 34.9	71.3 ± 16.9	86.7 ± 21.9	77.7 ± 22.7	80.0 ± 42.1	75.6 ± 15.7	83.7 ± 9.90	88.3 ± 13.8
LR	91.7 ± 18.0	66.7 ± 45.9	93.4 ± 10.8	85.7 ± 24.9	88.7 ± 15.7	63.7 ± 51.6	**95.1 ± 12.1**	81,5 ± 19.8	90.6 ± 19.4
C4.5	33.3 ± 24.1	26.7 ± 35.4	55.1 ± 15.8	66.7 ± 27.7	41.6 ± 14.1	63.7 ± 47.4	59.8 ± 10.2	59.3 ± 18.6	69.8 ± 11.1
K-NN	83.3 ± 24.1	53.3 ± 40.8	72.9 ± 14.8	66.7 ± 27.7	58.5 ± 16.3	45.4 ± 49.7	63.4 ± 19.9	68.5 ± 9.30	71.7 ± 8.80

NB (naive Bayes), K2 (K2 algorithm), LR (logistic regression), C4.5 (C4.5 algorithm), K-NN (k-nearest neighbor algorithm)

In response to accuracy concerns in the global models, new specific models were developed. Each model is associated with one of the nine journal sections, one of the four time horizons and one of the five supervised classification methods studied. Table 5.3 shows results for the new models. It shows that the results depend of the journal section, the time horizon and the supervised classification method used. The highest percentage of correctly classified cases is 100%, which was achieved on three occasions by the naive Bayes and logistic regression methods. On the other hand, the results were poorest for the C4.5 and K-NN methods that output values of less than 50%. Table 5.3 also shows the number of features accounted for the different predictive models. Fixing a specific journal section and analyzing the average number of features within the four-year time horizon, it is observed that sections with fewer features are *6-Phylogenetics* (18.75) and *2-Databases and Ontologies* (26.5), whereas the sections with most features are *3-Gene Expression* (112.75) and *7-Sequence Analysis* (97.5).

Looking at the behavior of the classifier for each value to be predicted {*few*, *some*, *many*}, Table 5.4 shows the confusion matrices of the models associated with sections *1-Data and Text Mining* and *2-Databases and Ontologies*, and with the logistic regression and decision trees methods, respectively. These models were chosen because they are the ones that are most and least accurate within each of the four time horizons, respectively (see Table 5.3).

To check the good behavior of the logistic regression method, the confusion matrix of *1-Data and Text Mining* and the *second-year* model is analyzed. This matrix shows that the total number of cases to be predicted is 88. Of these, 85 cases are well classified (96.6%) and three cases are wrongly classified (3.4%). Analyzing each value of the class, note that the success rate for the values *few* and *some* is 100%, whereas three errors are made for the value *many*, where one is classified as *few* and two as *some*. On the other hand, the confusion matrix of the *2-Databases and Ontologies* and *fourth-year* model is an example of the poor behavior of C4.5. In this case, the model tries to predict 15 cases, of which 4 are well classified (26.7%) and the rest are wrongly classified (73.3 %). Analyzing the different values of the class, it is observed that the success rate for the values for *few* and *many* is 33%, whereas

Table 5.4: Confusion matrices of two specific models (logistic regression and decision trees models).

Section *1-Data and Text Mining* (logistic regression)

First-year (98.9 ± 3.53)				*Second-year* (96.6 ± 5.43)				*Third-year* (94.0 ± 9.78)				*Fourth-year* (91.7 ± 18.0)			
a	b	c	← Classified as	a	b	c	← Classified as	a	b	c	← Classified as	a	b	c	← Classified as
81	0	0	a=few	**39**	0	0	a=few	**10**	0	0	a=few	**8**	0	0	a=few
1	**6**	0	b=some	0	**31**	0	b=some	0	**17**	0	b=some	0	**7**	0	b=some
0	0	**0**	c=many	1	2	**15**	c=many	0	3	**20**	c=many	2	0	**7**	c=many

Section *2-Databases and Ontologies* (C4.5)

First-year (86.5 ± 13.2)				*Second-year* (37.8 ± 11.3)				*Third-year* (48.0 ± 12.3)				*Fourth-year* (26.7 ± 35.4)			
a	b	c	← Classified as	a	b	c	← Classified as	a	b	c	← Classified as	a	b	c	← Classified as
32	0	0	a=few	**7**	5	1	a=few	**1**	0	5	a=few	**2**	0	4	a=few
3	**0**	0	b=some	7	**6**	1	b=some	2	**0**	6	b=some	0	**0**	3	b=some
2	0	**0**	c=many	6	3	**1**	c=many	0	0	**11**	c=many	2	2	**2**	c=many

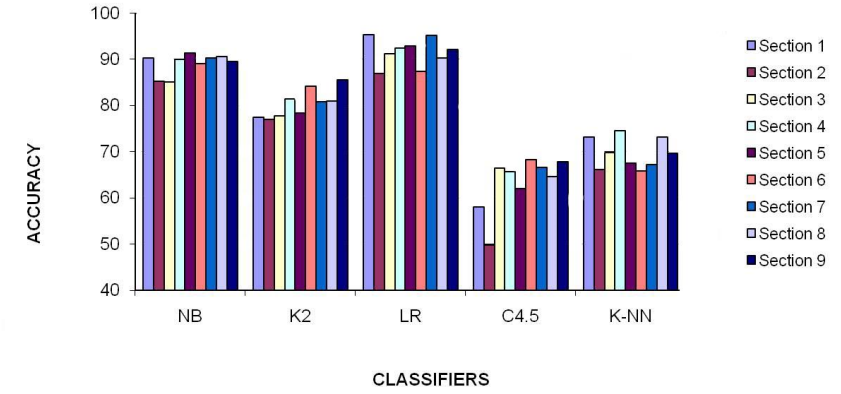


Figure 5.1: Average accuracy within the four prediction years by each section and classifiers.

success for the value for *some* is 0%, where all instances of this value are classified as *many* rather than as *some*.

The height of the bars in Figure 5.1 indicates the average percentages for the different classifiers within the four time horizons, with a fixed journal section. Taking the first bar as an example, the value displayed is 90.3%. This value is the mean accuracy for naive Bayes applied to *1-Data and Text Mining* averaged across the four time horizons. Figure 5.1 also shows that the journal section predicted with the highest success rate is method dependent. Logistic regression and naive Bayes achieve some notable results. Logistic regression predicts the *1-Data and Text Mining* journal section with a 95.30% success rate across the four time horizons, whereas naive Bayes predicts the *5-Genome Analysis* with an average accuracy of 91.32% across the four time horizons. On the other hand, the *4-Genetics and Population Analysis* section has the highest average percentage of cases well classified by all five algorithms (80.82%), whereas *2-Databases and Ontologies* is the section with the lowest percentage of well classified cases with an average accuracy of 73.05%.

Figure 5.2 indicates the average percentages scored by the different classifiers for the nine journal sections studied with a fixed year of publication. Taking the first bar as an example, the value displayed is 95.70%. This value is the mean accuracy of applying the naive Bayes classification method for the *first-year* of publication averaged over all journal sections. The best average results are for the *first* time horizon at 92.06% across all classifiers. The second, third and fourth time horizons have many similarities with each other, where percentages range from 73% to 75%. Looking at the scores for each algorithm, note that naive Bayes, K2, C4.5 and K-NN predict the *first-year* more accurately. However, logistic regression predicts the *third-year* more accurately, although this result is not significant compared with *first-* and *second-year* results (Figure 5.1).

After analyzing all results, it is observed that logistic regression and naive Bayes are the methods that solve the problem more accurately. Comparing these methods, logistic regression achieves a higher success rate, scoring 91.55% on average across the nine sections

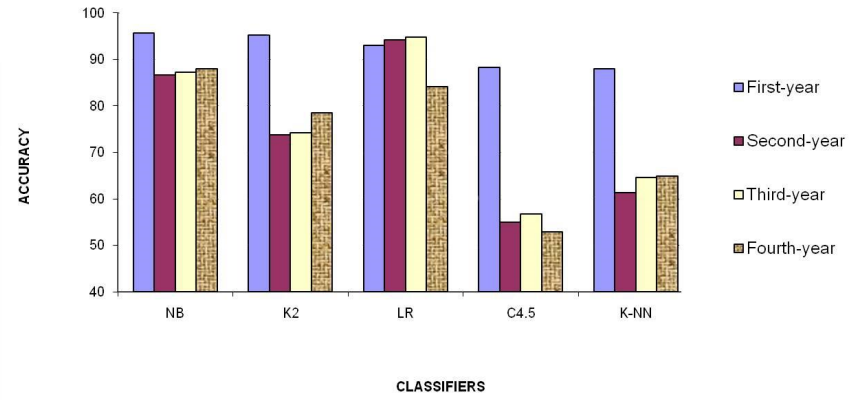


Figure 5.2: Average accuracy within the nine journal sections by each prediction year and classifier.

within the four time horizons, whereas naive Bayes attains 89.38%. Additionally, these methods are the only ones that correctly classified 100% of cases for a specific year and section (Table 5.3). Regarding the journal sections and time horizons, logistic regression specializes in section *1-Data and Text Mining* (95.30%) and in the *third-year* (94.78%), whereas naive Bayes specializes in *5-Genome Analysis* (91.32%) and the *first-year* (95.70%).

5.2.4 Exploiting the best models

The purpose of this section is to find out whether there are any tokens that influence an article's citation counts within the journal sections and time horizons. This analysis shows the results of predicting the number of citations of a new article using the models of the journal section *6-Phylogenetics* in the *third-year* learned by naive Bayes and logistic regression models (18 features, see Table 5.3).

Analyzing the probability distributions stored in the features of the naive Bayes model, it is observed the fact that an article receives *few*, *some* or *many* citations determines the probability of occurrence of tokens in the article. Similarly, if some tokens appear in an article, they influence the citation count, and thus determine the value of the class to be predicted. The three probability columns $P(X_i|f)$, $P(X_i|s)$ and $P(X_i|m)$ in Table 5.5 show the distributions of each token subject to the class values. These distributions show that there are some tokens like *linear*, *probability*, *discussed*, *automated*, *time* and *nucleotide*, which tend to appear more frequently in the papers with *few* citations. On the other hand, for papers that have *some* citations, these tokens are *time*, *nucleotide*, *dynamic*, *entire*, *independent*, *compared*, *interaction*, *clustering* and *protein*. Finally, *parameter*, *performance*, *analyze*, *researchers*, *likelihood based*, *linear* and *probability* are the tokens with a higher frequency of occurrence in articles that receive *many* citations. The above probabilities and the marginal probability of each class value $P(C{=}c)$ with $c = f,\ s,\ m$ (Table 5.1), are the elements of the naive Bayes model used to predict the citation count of a specific paper ($\boldsymbol{x}$). This model is:

$$P(C = c \mid \boldsymbol{x}) \propto P(C = c) \prod_{i=1}^{n} P(X_i = x_i \mid C = c) \tag{5.1}$$

On the other hand, the logistic regression model requires some coefficients (β_i) to calculate the class value with higher a posteriori probability. These coefficients are shown in the middle columns (β_i^f and β_i^s) of Table 5.5. The models used for these predictions are:

$$P(C = f \mid \boldsymbol{x}) = \frac{e^{(\beta_0^f + \sum_{i=1}^{n} \beta_i^f x_i)}}{1 + e^{(\beta_0^f + \sum_{i=1}^{n} \beta_i^f x_i)} + e^{(\beta_0^s + \sum_{i=1}^{n} \beta_i^s x_i)}} \tag{5.2}$$

$$P(C = s \mid \boldsymbol{x}) = \frac{e^{(\beta_0^s + \sum_{i=1}^{n} \beta_i^s x_i)}}{1 + e^{(\beta_0^f + \sum_{i=1}^{n} \beta_i^f x_i)} + e^{(\beta_0^s + \sum_{i=1}^{n} \beta_i^s x_i)}} \tag{5.3}$$

$$P(C = m \mid \boldsymbol{x}) = 1 - P(C = f \mid \boldsymbol{x}) - P(C = s \mid \boldsymbol{x}) \tag{5.4}$$

The new case to be predicted is shown in the last column of Table 5.5. This new case is a paper abstract. *Analyze*, *researchers*, *automated*, *nucleotide*, *dynamic*, *entire*, *compared* and *clustering* are the tokens that appear in the abstract. After propagating this evidence, the results predicted by naive Bayes are $P(f|\boldsymbol{x}) = 0.30$, $P(s|\boldsymbol{x}) = 0.67$ and $P(m|\boldsymbol{x}) = 0.03$. On the other hand, the results predicted by logistic regression are $P(f|\boldsymbol{x}) = 0.18$, $P(s|\boldsymbol{x}) = 0.81$ and $P(m|\boldsymbol{x}) = 0.01$. The results of both models show that an abstract with the above tokens published in the journal section *6-Phylogenetics* will receive *some* citations (i.e. 2, 3, or 4 citations) in the *third-year* after publication.

Table 5.5: Exploiting the best models. Naive Bayes and logistic regression predictions of the number of citations in the third year of a new article published in section *6-Phylogenetics*.

Feature (X_i)	$P(X_i = 1 \mid C = c)$			*Coeff. LR*		*New article*
	$P(X_i \mid f)$	$P(X_i \mid s)$	$P(X_i \mid m)$	β_i^f	β_i^s	$\boldsymbol{x}$
parameter	0.25	0.09	**0.27**	-6.76	-10.93	
performance	0.25	0.09	**0.27**	-6.76	-10.93	
analyze	0.25	0.09	**0.36**	-12.35	-11.57	√
researchers	0.25	0.09	**0.27**	-7.53	-10.93	√
likelihood based	0.25	0.09	**0.27**	-8.13	-10.93	
linear	**0.50**	0.09	**0.27**	-13.17	-11.57	
probability	**0.50**	0.09	**0.27**	-3.18	-11.57	
discussed	**0.75**	0.09	0.09	28.34	-10.93	
automated	**0.50**	0.09	0.09	26.85	-10.35	√
time	**0.50**	**0.36**	0.09	12.38	7.22	
nucleotide	**0.50**	**0.36**	0.09	12.38	7.22	√
dynamic	0.25	**0.27**	0.09	5.22	12.19	√
entire	0.25	**0.27**	0.09	5.22	12.20	√
independent	0.25	**0.36**	0.09	5.53	12.91	
compared	0.25	**0.45**	0.09	5.88	13.72	√
interaction	0.25	**0.27**	0.09	5.22	12.20	
clustering	0.25	**0.27**	0.09	5.22	17.42	√
protein	0.25	**0.45**	0.09	5.88	13.72	
Intercept (β_0)				-16.1833	-3.6972	

5.3 Discussion and conclusions

Nowadays, publishers of scientific journals face the tough task of selecting high quality articles that will attract as many readers as possible from a pool of articles. The possibility of a journal having a tool capable of predicting the citation count of an article within the first few years after publication would pave the way for new assessment systems.

This chapter presents a new approach based on building several prediction models for the *Bioinformatics* journal. These models predict the citation count of an article within four years after publication (global models). To build these models, tokens found in the abstracts of *Bioinformatics* papers have been used as predictive features, along with other features like the journal sections and two-week post publication periods. To improve the accuracy of the global models, specific models have been built for each *Bioinformatics* journal section (*Data and Text Mining, Databases and Ontologies, Gene Expression, Genetics and Population Analysis, Genome Analysis, Phylogenetics, Sequence Analysis, Structural Bioinformatics and Systems Biology*).

Results of specific models achieved a greater rate of success across the four years than the global models. The logistic regression and naive Bayes classification methods output high average scores in the nine journal sections and across the four time horizons, achieving rates of 91.5% (AUC=0.943) and 89.4% (AUC=0.983) respectively. It is also observed that the appearance of certain words in the paper abstracts can influence the number of citations received. The probabilities assigned and the tokens selected depend on the journal section and chosen time horizon. The selected tokens could be used as a point of reference to identify the hot topics.

Unlike previous models [64, 80, 296], the predictions of these models are not based on information available after publication. The proposed models use the information content of the article abstract. In this way, predictions can be made at publication time, and it is not necessary to wait until the end of a data collection period to predict citation count. It could be worthwhile comparing the proposed models with models developed by Fu and Aliferis [164] because, although they use different features, data sets, response variable and prediction horizon, they both attempt to predict citations before publication with tokens contained in the article abstract. However, the estimated accuracy of the naive Bayes (AUC=0.983) and logistic regression (AUC=0.943) supervised classification methods were higher than the accuracy achieved by their models (AUC=0.918).

In the future, the target will be to build new models that incorporate other paper-based features (title, keywords, conclusions, etc.), new author-based features (h-index, number of papers, number of citations, etc.) and new journal-based features (impact factor, immediacy index, category, etc.). These models would be induced using different machine learning methods. The way citation count is handled influences the results. It could be modeled as a continuous variable using other methods like regression, regularized regression, or local regression.

Chapter 6

Predicting the h-index using cost-sensitive algorithms

6.1 Introduction

Classification problems commonly assume that the class values are unordered. But, these values have a natural order in many practical applications. Given ordered classes, it is not only interested in maximizing classification accuracy, but also in minimizing the distances between the actual and the predicted classes. Some fields, like statistics, have faced this problem for many years, developing several approaches [313, 314], whereas other fields, like machine learning, have only recently started to look at the problem [79, 105, 162, 166, 265, 292, 365, 412].

Researchers have tried to solve the above problem by means of different approaches. For example, Kramer *et al.* [267] transformed ordinal scales into numeric values, and then solve the problem as a standard regression problem. Also, Frank and Hall [161] used binary decomposition techniques, transforming the original problem involving k classes into k-1 binary problems. The cost-sensitive learning approach is also used for the above purpose. Direct cost-sensitive algorithms design classifiers that directly use misclassification costs in the learning algorithms [128, 294, 442]. In contrast, indirect cost-sensitive algorithms convert existing cost-insensitive classifiers into cost-sensitive classifiers [126, 414, 436, 478, 487].

This chapter incorporates the direct cost-sensitive learning and feature subset selection into the well-known naive Bayes [325], which is the most straightforward and widely tested method for probabilistic induction and has long been used within the field of pattern recognition [129]. New cost-sensitive algorithms based on the selective naive Bayes notions [275] are developed. These direct algorithms add misclassification costs to the learning algorithm, and use wrapper approaches to select relevant variables that maximize the accuracy (*CS-SNB-Accuracy* algorithm) and minimize the cost (*CS-SNB-Cost* algorithm). The objective of these approaches is to build parsimonious models. These models will not include features that are irrelevant and redundant. Some benefits of applying variable selection are better

classification performance, faster classification models, smaller databases, and the ability to gain more insight into the process that is being modeled.

Only a few approaches tackled Bayesian classifiers using cost-sensitive approaches. For example, Gama [167] presented a cost-sensitive iterative Bayes. For another example, Chai *et al.* [83] specifically consider test-cost sensitive learning and propose a test-cost sensitive naive Bayes. For the third example, Fang [148] develops a cost-sensitive naive Bayes method which learns and infers the order relation from the training data and classifies the instance based on the inferred order relation. Finally, Jiang *et al.* [241] incorporated an indirect cost-sensitive method, called instance weighted, into naive Bayes, tree augmented Bayesian networks, averaged one dependence estimators and hidden naive Bayes to make these Bayesian network classifiers cost-sensitive.

The interest and originality of this chapter is two-fold. First, two new classifiers (*CS-SNB-Accuracy* and *CS-SNB-Cost*) are proposed to bring together the advantages of using the cost-sensitive learning approach and the feature subset selection. Second, both classifiers have been tested on the bibliometric indices prediction area, that is, they are used to predict the annual increase of the *h-index* for scientific journals belonging to the Journal Citation Report Neurosciences category across a four-year time horizon using bibliometric indices. This chapter is based on the published paper [221].

Chapter outline

The remainder of this chapter is organized as follows. Section 6.2 explains some concepts related to cost-sensitive Bayesian classifiers, focusing on the proposed new cost-sensitive selective naive Bayes approaches. Section 6.3 presents the main results, including dataset construction, data distribution, accuracy and average cost of predictive models and some examples. Finally, Section 6.4 outlines some conclusions emphasizing the original contribution of the paper and future research on the topic.

6.2 Cost-sensitive Bayesian classifiers

The objective of cost-sensitive methods is to take into account misclassification costs different from 0 (hit) and 1 (miss). These methods are concerned with classification accuracy and classification costs. Two forward cost-sensitive selective naive Bayes approaches are developed. The search process of the first approach (*CS-SNB-Accuracy*) is based on maximizing classification accuracy, that is, it includes variables that improve classification accuracy, whereas the search process of the second approach (*CS-SNB-Cost*) is based on minimizing misclassification costs, that is, it includes variables that reduce the distances between the actual and the predicted classes.

Given a cost matrix and a set of predicted class probabilities for each instance, both approaches readjust the probability thresholds of each class to select the class with the minimum-expected cost. The expected cost of each prediction is obtained by multiplying

the associated costs by the predicted class probabilities. Unlike selective naive Bayes, these approaches do not select the most likely class value of the posterior distribution, they select the class (c^*) that minimizes the expected cost of predictions given a new instance $\mathbf{x}$:

$$c^* = \arg \min_{c \in \Omega(C)} \sum_{c' \in \Omega(C)} p(c' \mid \mathbf{x}) \;\; cost(c \mid c') \tag{6.1}$$

where

$$p(c' \mid \mathbf{x}) \propto p(c') \prod_{i=1}^{m} p(x_i \mid c') \prod_{j=m+1}^{n} \mathcal{N}(x_j, \mu_{c'j}, \sigma^2_{c'j}) \tag{6.2}$$

and $cost(c \mid c')$ is the associated misclassification cost.

In short, the first approach (*CS-SNB-Accuracy*) considers adding each variable to the model and measures the performance of the resulting model on the training data. The variable that most improves the accuracy, that is, the percentage of well-classified instances in the predicted class (c^*) over the actual class, is permanently added to the model. In contrast, the second approach (*CS-SNB-Cost*) considers adding variables that reduce the misclassification cost between the predicted and actual class.

6.2.1 Cost-sensitive selective naive Bayes - Accuracy

Algorithm 1 shows the pseudocode of the cost-sensitive selective naive Bayes - Accuracy model. This algorithm chooses k-fold cross-validation as the procedure for estimating the accuracy and cost of models classifying new cases according to the value of the predictive features. This method is stratified, that is, it divides all cases into k disjoint subsets of approximately equal proportion of class values and equal size. Each subset is used to test a model that is learned from the other *k-1* subsets.

The *trainingSetGeneration* and *testSetGeneration* functions provide the required subsets of cases in each iteration. This algorithm initializes the model to the class variable, that is, there is no predictive variables in the model yet. After that, the algorithm saves the accuracy of the resulting model (*theresholdAccuracy*) for subsequent comparisons. The accuracy threshold is computed by means of *estimateClassProb* and *max* functions, which, respectively, compute the initial class probabilities given the training set and return the highest probability value, that is, the probability of the most frequent class. In each iteration, the algorithm checks if a specific variable belongs to the model. The *isModelVariable* function returns true or false according to the current model. The algorithm considers adding each unused variable to the model on a trial basis and measures the performance of the resulting model on the training data. First, the *predictClass* function computes the predicted class, that is, the most likely class value of the posterior distribution given a case of the training set. Then, the *readjustClass* function readjusts the probability thresholds of each class to select the class with the minimum-expected cost (see Equation 6.1). Finally, the readjusted class and the actual class are used to calculate the model's accuracy (*calculateAccuracy*) using

```
Input  : Dataset (feature variables and class variable) and cost matrix
Output: Accuracy and cost of the cost-sensitive selective naive Bayes - Accuracy model
for k ← 1 to folds do
    // k-fold cross validation
    trainingSet ← trainingSetGeneration(dataset,k);
    testSet ← testSetGeneration(dataset,k);
    // Training phase
    model ← {class};
    initialProbability ← estimateClassProb(trainingSet,model);
    thresholdAccuracy ← max(initialProbability);
    accuracyVector ← {};
    continue ← true;
    while continue do
        for variable ← 1 to size(numVariables) do
            sw ← isModelVariable(variable);
            if sw then
                for case ← 1 to size(trainingSet) do
                    actualClass ← getActualClass(case);
                    predictedClass ← predictClass (case,model);
                    readjustedClass ← readjustClass (predictedClass, costMatrix);
                    if (actualClass==readjustedClass) then hit=1;
                    else hit=0 modelAccuracy ← calculateAccuracy(hit)
                end
            end
            accuracyVector[variable] ← modelAccuracy
        end
        accuracy ← max(accuracyVector);
        if (accuracy > thresholdAccuracy) then
            bestVariable ← selectBestVariable(accuracyVector);
            model ← addToModel(model, bestVariable);
            thresholdAccuracy ← accuracy;
        else
            continue ← false;
        end
    end
    // Test phase
    for case ← 1 to size(testSet) do
        actualClass ← getActualClass(testSet, case);
        predictedClass ← predictClass (testSet, model);
        readjustedClass ← readjustClass (predictedClass, costMatrix);
        if (actualClass==readjustedClass) then hit=1;
        else hit=0 accuracy ← calculateAccuracy(hit);
        cost ← calculateCost(actualClass,readjustedClass);
    end
    finalAccuracyVector ← addToVector(accuracy);
    finalCostVector ← addToVector(cost);
end
finalAccuracy ← mean(finalAccuracyVector);
finalCost ← mean(finalCostVector);
```

Algorithm 1: Cost-sensitive selective naive Bayes - Accuracy model

the selected unused variable in each iteration. After computing the models' accuracies of all unused variables, the best variable (*selectBestVariable*), that is, the variable related to the model with the highest accuracy is preselected to be added to the final model. If the new accuracy is higher than the current accuracy threshold, then the variable is permanently added to the final model (*addToModel*). The algorithm terminates when the addition of any variable results in reduced accuracy. During the test phase, the algorithm computes the accuracy (*calculateAccuracy*) and cost (*calculateCost*) of the model classifying the cases belonging to the test set. Finally, the k percentages of well-classified cases and the k misclassification costs are averaged to output the estimated values of the model learned from all cases to classify new cases.

6.2.2 Cost-sensitive selective naive Bayes - Cost

Algorithm 2 shows the pseudocode of the cost-sensitive selective naive Bayes - Cost model. This algorithm also chooses k-fold cross-validation as the procedure for estimating the accuracy and cost of the models.

This algorithm initializes the model to the class variable, that is, there is not predictive variables in the model yet. After that, the algorithm saves the misclassification cost of the resulting model (*theresholdCost*) for subsequent comparisons. The cost threshold is computed by means of *estimateClassProb*, *classCostEstimation* and *min* functions, which, respectively, compute the initial class probabilities given the training set and the model, compute the initial cost given the initial probabilities and the cost matrix, and finally, return the lowest cost value. In each iteration, the algorithm checks if a specific variable belongs to the model. The *isModelVariable* function returns true or false according to the current model. The algorithm considers adding each unused variable to the model on a trial basis and measures the average cost of the resulting model on the training data. First, the *predictClass* function computes the predicted class, that is, the most likely class value of the posterior distribution given a case of the training set. Then, the *readjustClass* function readjusts the probability thresholds of each class to select the class with the minimum-expected cost (see Equation 6.1). Finally, the readjusted class and the actual class are used to calculate the model's cost (*calculateCost*) using the selected unused variable in each iteration. After computing the costs of all models, the best variable (*selectBestVariable*), that is, the variable associated with the model with lowest misclassification cost is preselected to be added to the final model. If the new model's cost is lower than the current cost threshold, then the variable is permanently added to the model (*addToModel*). The algorithm terminates when the addition of any variable results in a higher cost. During the test phase, the algorithm computes the accuracy (*calculateAccuracy*) and cost (*calculateCost*) of the model classifying the cases belonging to the test set. Finally, the k percentages of well-classified cases and the k misclassification costs are averaged to output the estimated values of the model learned from all cases to classify new cases.

```
Input  : Dataset (feature variables and class variable) and cost matrix
Output: Accuracy and cost of the cost-sensitive selective naive Bayes - Cost model
for k ← 1 to folds do
    // k-fold cross validation
    trainingSet ← trainingSetGeneration(Dataset,k);
    testSet ← testSetGeneration(Dataset,k);
    // Training phase
    model ← {class};
    initialProbability ← estimateClassProb(trainingSet,model);
    initialCost ← classCostEstimation(initialProbability,costMatrix);
    thresholdCost ← min(initialCost);
    accuracyVector ← {};
    continue ← true;
    while continue do
        for variable ← 1 to size(numVariables) do
            sw ← isModelVariable(variable);
            if sw then
                for case ← 1 to size(trainingSet) do
                    actualClass ← getActualClass(case);
                    predictedClass ← predictClass (case,model);
                    readjustedClass ← readjustClass (predictedClass,costMatrix);
                    modelCost ← calculateCost(actualClass,readjustedClass)
                end
            end
            costVector[variable] ← modelCost
        end
        cost ← min(costVector);
        if (cost < thresholdCost) then
            bestVariable ← selectBestVariable(costVector);
            model ← addToModel(model, bestVariable);
            thresholdCost ← cost;
        else
            continue ← false;
        end
    end
    // Test phase
    for case ← 1 to size(testSet) do
        actualClass ← getActualClass(testSet, case);
        predictedClass ← predictClass (testSet, model);
        readjustedClass ← readjustClass (predictedClass, costMatrix);
        if (actualClass==readjustedClass) then hit=1;
        else hit=0 accuracy ← calculateAccuracy(hit);
        cost ← calculateCost(actualClass,readjustedClass);
    end
    finalAccuracyVector ← addToVector(accuracy);
    finalCostVector ← addToVector(cost);
end
finalAccuracy ← mean(finalAccuracyVector);
finalCost ← mean(finalCostVector);
```

Algorithm 2: Cost-sensitive selective naive Bayes - Cost model

6.3 Predicting the h-index of Neuroscience journals

6.3.1 Dataset compilation

Web of Science and Journal Citation Reports are selected as sources to download publication and citation data. First, all journals belonging to the Journal Citation Reports Neurosciences category from 2000 to 2011 are selected. There were 269 journals in this category during the analyzed period. Then the publication list and citation data were obtained for these journals from the Web of Science. All documents (1,044,811 papers) published by the 269 journals were downloaded until 2011. Using the above information, some scientific impact indices (*documents, citations, the h-index, the g-index, the hg-index, the a-index, the m-index, the* q^2*-index, the* h_r*-index, the* h_i*-index* and *the* h_c*-index*) were calculated for each journal from 2000 to 2011. Furthermore, other specific journal indices values (*impact factor, immediacy index, cited half-life, eigenfactor* and *article influence*) were also downloaded from Journal Citation Reports. Finally, all information was stored in a database designed for this purpose.

6.3.2 Data distribution

After collecting the publication list and citation data of all journals, it is observed that the number of cases selected to build the predictive models varied depending on the year. We used journal data from 2000 to 2010 (2305 cases) to construct the models assigned to the first-year. On the other hand, the models for the second-year used journal data from 2000 to 2009 (2037 cases). Finally, the predictive models for the third- and fourth-year used journal data from 2000 to 2008 (1785 cases) and from 2000 to 2007 (1449 cases), respectively. Clearly, the longer the prediction horizon was the fewer cases were used to induce the models.

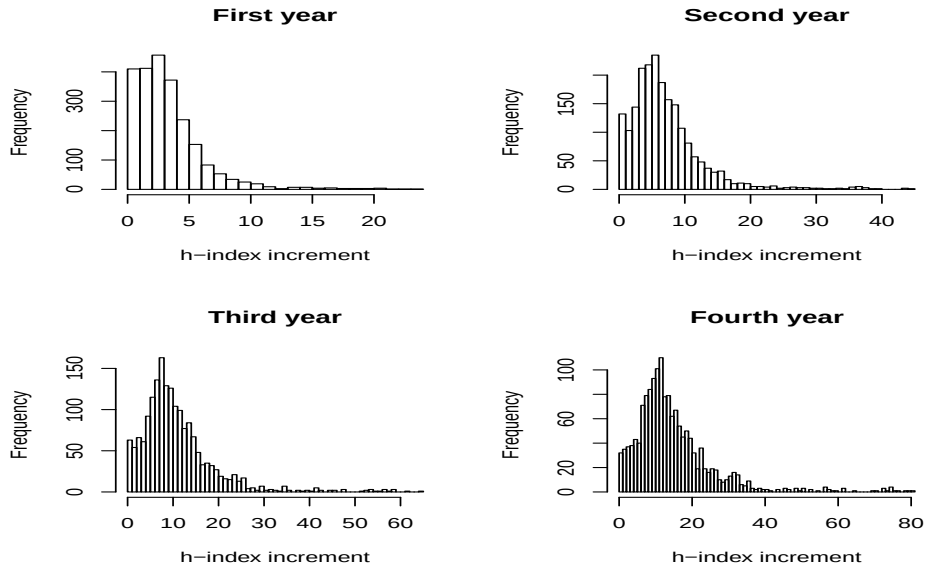


Figure 6.1: Distribution of the increase of the h-index for different prediction years

Figure 6.1 shows the distribution of the journals selected according to the annual increase of their *h-index* value within the first four years. Taking the first year as an example, it is observed that the lowest and highest increment of the *h-index* was Δh=0 (128 journals) and Δh=24 (1 journal), respectively. Note that 457 journals had an increase of Δh=3, which was the mode value for the first year. Regarding the second year, the minimum value was Δh=0 (42 journals), the maximum value was Δh=45 (1 journal) and the mode value was Δh=6 (235 journals). Finally, it is also noted that the *h-index* value increased from Δh=0 to Δh=65 for third-year models and from Δh=0 to Δh=81 for fourth-year models. Their mode values were Δh=8 (163 journals) and Δh=12 (110 journals), respectively.

The class variable values were discretized into four intervals with equal frequency. The increment of the *h-index* values were assigned to one of the four possible class values (low, medium-low, medium-high and high). In this way, first-year models were discretized as low (Δh=[0-1]), medium-low (Δh=[2]), medium-high (Δh=[3-4]) and high (Δh=[$\geq$5]), whereas fourth-year models were discretized as low (Δh=[0-8]), medium-low (Δh=[9-12]), medium-high (Δh=[13-18]) and high (Δh=[$\geq$19]). The correspondence between Δh values and class labels for all models are shown in Table 6.1.

6.3.3 Predictive models

This section compared the proposed approaches with the standard formulation of selective naive Bayes in order to determine if their accuracy and average cost values were reasonable. Table 6.2 shows the estimated accuracy and the average cost for each model. Numbers in boldface represent the highest accuracy value and lowest cost value for each model.

The proposed methods were tested with different cost matrices ($C(0, n)$, $C(0, n^2)$, $C(0, 2^n)$ and $C(0, n^n)$). The cost matrix $C(0, n)$ represents costs where the correct classification has no costs and the incorrect classification has linear costs. Similarly, $C(0, n^2)$, $C(0, 2^n)$ and $C(0, n^n)$ represents costs where the correct classification has no costs and the incorrect classification has quadratic and exponential costs. Using the cost matrix $C(0, n^n)$, for example, it is observed that the proposed models almost always outperform the selective naive Bayes models in higher accuracy and lower cost. Although these models achieved the highest accuracy (0.504) in the first year, the proposed two new algorithms, specifically the *CS-SNB-Accuracy*, achieved the highest accuracy in the second-year (0.518), third-year (0.542) and fourth-year (0.532). By average cost, it is observed that the proposed models, specifically

Table 6.1: Correspondence between Δh values and class labels (low, medium-low, medium-high and high) after discretization with equal frequency

	First-year	Second-year	Third-year	Fourth-year
Low values	0-1	0-4	0-6	0-8
Medium-Low values	2	5-6	7-9	9-12
Medium-High values	3-4	7-9	10-14	13-18
High values	$\geq$5	$\geq$10	$\geq$15	$\geq$19

Table 6.2: Accuracy and average cost of models which are learned using different selective naive Bayes approaches and cost matrices

	First year		Second year		Third year		Fourth year	
Methods	Accur	Cost	Accur	Cost	Accur	Cost	Accur	Cost
Cost matrix: $C(0, n)$								
Selective naive Bayes	**0.502**	0.608	**0.506**	0.644	0.530	0.563	0.517	0.584
CS-SNB-Accuracy	0.477	0.610	0.513	0.579	0.530	0.543	0.528	**0.534**
CS-SNB-Cost	0.458	**0.596**	0.519	**0.577**	**0.534**	**0.538**	**0.532**	0.546
Cost matrix: $C(0, n^2)$								
Selective naive Bayes	**0.503**	0.828	0.501	1.005	0.525	0.758	0.532	0.742
CS-SNB-Accuracy	0.460	**0.721**	0.498	0.873	0.515	0.766	0.525	0.713
CS-SNB-Cost	0.451	0.735	**0.514**	**0.775**	**0.532**	**0.708**	**0.533**	**0.706**
Cost matrix: $C(0, 2^n)$								
Selective naive Bayes	**0.507**	1.211	0.509	1.299	0.514	1.171	0.518	1.170
CS-SNB-Accuracy	0.480	1.221	**0.519**	**1.156**	**0.538**	**1.069**	0.523	1.101
CS-SNB-Cost	0.460	**1.187**	0.507	1.168	0.533	1.080	**0.532**	**1.086**
Cost matrix: $C(0, n^n)$								
Selective naive Bayes	**0.504**	1.227	0.506	1.327	0.526	1.133	0.516	1.190
CS-SNB-Accuracy	0.446	0.953	**0.518**	**0.769**	**0.542**	0.714	**0.532**	0.732
CS-SNB-Cost	0.419	**0.753**	0.500	0.772	0.513	**0.695**	0.516	**0.705**

the *CS-SNB-Cost*, always achieve a lower cost than the selective naive Bayes. Taking the first-year as an example, it is found that cost associated with the selective naive Bayes is 1.227, whereas the cost related to the *CS-SNB-Cost* is 0.753.

Focusing on cost matrices, it is observed that accuracy varied across models and prediction years, but a general pattern were not found. The selective naive Bayes model achieved the highest accuracy value for first year (0.507) using the cost matrix $C(0, 2^n)$. In contrast, the *CS-SNB-Accuracy* model obtained the highest accuracy value in the second year (0.519) and third year (0.542) with the cost matrix $C(0, 2^n)$ and $C(0, n^n)$, respectively. Finally, *CS-SNB-Cost* achieved the highest accuracy values in the fourth year (0.533) using the cost matrix $C(0, n^2)$. By costs, it is observed that the lowest and highest average cost were achieved by $C(0, n)$ and $C(0, 2^n)$. *CS-SNB-Cost* almost always achieved the lowest average cost with the $C(0, n)$ matrix.

Regarding each algorithm, note that selective naive Bayes achieved the highest accuracy values for first-year models no matter which cost matrix was used. In contrast, *CS-SNB-Accuracy* and *CS-SNB-Cost* predicted almost all the values more accurately than selective naive Bayes for the other prediction years. Given the cost matrix $C(0, 2^n)$, for example, it is found that selective naive Bayes achieved the highest accuracy (0.507) for first-year models, *CS-SNB-Accuracy* achieved the highest accuracy for second- (0.519) and third-year (0.538) models, and *CS-SNB-Cost* achieved the highest accuracy (0.532) for fourth-year models. Analyzing average cost, it is observed that the selective naive Bayes value was never the lowest. The lowest average cost values were always achieved by cost-sensitive models. Specially, note that *CS-SNB-Cost* usually obtained the lowest value. Given the cost matrix $C(0, n^2)$, for example, we noted that *CS-SNB-Accuracy* achieved the lowest average cost (0.721) for first-year

Table 6.3: Accuracy and average cost of models which are learned using different classification methods. Results are achieved using the cost matrix C(0, n^n) for all prediction years

	First year		Second year		Third year		Fourth year	
Methods	Accur	Cost	Accur	Cost	Accur	Cost	Accur	Cost
NB	0.262†	3.138†	0.343†	2.815†	0.306†	2.718†	0.303†	2.779†
SNB	0.504†	1.227†	0.506†	1.327†	0.526†	1.133†	0.516†	1.190†
CS-SNB-Accuracy	0.446†	0.953†	0.518†	**0.769**	0.542†	0.714	0.532†	0.732
CS-SNB-Cost	0.419†	**0.753**	0.500†	0.772	0.513†	**0.695**	0.516†	**0.705**
C4.5	0.525†	1.204†	0.598	1.073†	0.640	0.793†	0.654	0.879†
K-NN	0.553†	1.113†	**0.609**	1.080†	**0.643**	0.803†	**0.655**	0.857†
Logistic	**0.587**	0.978†	0.587†	1.058†	0.622	0.866†	0.625	0.878†

Naive Bayes (NB); Selective naive Bayes (SNB); C4.5 decision tree (C4.5);
K-nearest neighbour (K-NN); Logistic regression (Logistic)

models, whereas *CS-SNB-Cost* achieved the lowest values for second- (0.775), third- (0.708) and fourth-year (0.706) models. To summarize, we found that our cost-sensitive approaches, particularly *CS-SNB-Cost*, almost always achieved a lower average cost than selective naive Bayes. Also, our approaches, specially *CS-SNB-Accuracy*, often obtained higher accuracy values than selective naive Bayes.

In order to compare the performance of the proposed algorithms, Table 6.3 shows the accuracy and average cost of a set of classifiers (naive Bayes , selective naive Bayes, cost-sensitive selective naive Bayes-Accuracy, cost-sensitive selective naive Bayes-Cost, C4.5 decision tree, K-nearest neighbour and logistic regression) which are learned using the cost matrix $C(0, n^n)$ for all prediction years. Results are evaluated using two non-parametric tests such as Kruskal-Wallis test and Mann-Whitney test which analyze whether samples could have come from the same distribution. The significance level of these tests was 0.05 in all cases.

Analyzing the accuracy values, it is distinguished three different groups (low, medium and high values). The first group is composed by the naive Bayes classifier which achieved the lower values. In contrast, the second group is composed by three classifiers (selective naive Bayes, cost-sensitive selective naive Bayes-Accuracy, and cost-sensitive selective naive Bayes-Cost) that achieve medium values, whereas the third group is composed by non Bayesian classifiers (C4.5 decision tree, k-nearest neighbour and logistic regression), having the highest values. Note the above behavior no matter which prediction year was used. The results of the Kruskal-Wallis test showed that there were significant differences among the seven classifiers on the basis of the accuracy. So, Mann-Whitney tests were run in order to find out which classifiers rank better according to this criterion. It is compared the benchmark classifier, which had the highest average value, with the other classifiers. Classifiers marked in Table 6.3 with the symbol † had statistically significant differences with respect to the benchmark classifier (highlighted in boldface). Taking the second-year model as an example, results show that there were significant differences between k-nearest neighbour (benchmark classifier) and naive Bayes, selective naive Bayes, *CS-SNB-Accuracy*, *CS-SNB-Cost* and logistic regression. In contrast, results do not show statistically significant differences between k-nearest neighbour and C4.5 decision tree.

Table 6.4: Accuracy and average cost of models which are learned using different cost-sensitive approaches. Values achieved using the cost matrix $C(0, 2^n)$ for all prediction years

	First year		Second year		Third year		Fourth year	
Methods	Accur	Cost	Accur	Cost	Accur	Cost	Accur	Cost
MetaCost	0.116†	1.770†	0.315†	1.597†	0.326†	1.347†	0.188†	1.632†
CostSensitiveClassifier	0.253†	1.866†	0.329†	2.382†	0.300†	2.005†	0.238†	1.842†
CSRoulette	0.432†	2.878†	0.517	2.923†	0.521†	2.923†	0.526	2.892†
CS-SNB-Accuracy	**0.480**	1.221†	**0.519**	**1.156**	**0.538**	**1.069**	0.523	1.101
CS-SNB-Cost	0.460†	**1.187**	0.507†	1.168	0.533	1.080	**0.532**	**1.086**

Regarding the cost values, it is also differentiated three groups. In this case, naive Bayes and selective naive Bayes achieved higher costs, whereas C4.5 decision tree, k-nearest neighbour and logistic regression achieved medium costs. Finally, the proposed classifiers, *CS-SNB-Accuracy* and *CS-SNB-Cost* achieved the lowest costs. A Kruskal-Wallis test was also performed in order to compare classifiers on the basis of the average cost. Taking the second-year model as an example, results show that there were significant differences between *CS-SNB-Cost* and naive Bayes, selective naive Bayes, C4.5 decision tree, k-nearest neighbour and logistic regression. In contrast, results do not show statistically significant differences between the two proposed cost-sensitive algorithms.

Analyzing different cost-sensitive approaches, it is compared the proposed algorithms with other cost-sensitive algorithms like MetaCost, CostSensitiveClassifier and CSRoulette. These classifiers convert existing cost-insensitive classifiers (e.g. naive Bayes) into cost-sensitive ones. Table 6.4 shows the accuracy and average cost of the above classifiers which are learned using the cost matrix $C(0, 2^n)$ for all prediction years.

Focusing on accuracy and cost values, it is observed in Table 6.4 that the proposed models outperform other cost-sensitive classifiers no matter which prediction year was used. Taking the first-year as an example, it is noted that the *CS-SNB-Accuracy* achieved the highest accuracy (0.480) whereas the *CS-SNB-Cost* achieved the lowest cost (1.187). The results of the Kruskal-Wallis test showed that there were significant differences among the classifiers on the basis of accuracy and cost. So, Mann-Whitney tests were run in order to find out which classifiers rank better according to these criteria. It is compared the benchmark classifier, which had the best average value, with the other classifiers. Classifiers marked in Table 6.4 with the symbol † had statistically significant differences with respect to the benchmark classifier (highlighted in boldface). Results show that there were significant differences between *CS-SNB-Accuracy* (benchmark classifier) and MetaCost, CostSensitiveClassifier, CSRoulette and *CS-SNB-Cost* in terms of accuracy. Results also show that there were significant differences between *CS-SNB-Cost* (benchmark classifier) and MetaCost, CostSensitiveClassifier, CSRoulette and *CS-SNB-Accuracy* in terms of costs. To summarize, it is found that the proposed cost-sensitive approaches, particularly *CS-SNB-Cost*, achieved a lower average cost than other cost-sensitive classifiers. Also, our approaches, specially *CS-SNB-Accuracy*, obtained higher accuracy values than other cost-sensitive classifiers.

Let us now analyze the models in more detail. Table 6.5 shows the specific variables,

Table 6.5: Variables, accuracy and cost for the *CS-SNB-Cost* model by each fold of the cross-validation process. Values achieved using the cost matrix $C(0, n^2)$ for all prediction years

First year				Second year			
k	Variables	Accuracy	Cost	k	Variables	Accuracy	Cost
1	12,11	0.456	0.721	1	12,11,14	0.497	0.866
2	12,11	0.478	0.756	2	12,16,14	0.566	0.783
3	12,11	0.465	0.713	3	12,16,14	0.517	0.733
4	12,14	0.391	0.834	4	12,16,14	0.492	0.847
5	12,14	0.521	0.647	5	12,16	0.507	0.783
6	12,11	0.456	0.713	6	12,11,14	0.492	0.788
7	12,11	0.439	0.730	7	12,16,14	0.517	0.793
8	12,11	0.447	0.682	8	12,16,14	0.566	0.596
9	12,14	0.426	0.765	9	12,16	0.522	0.778
10	12,11	0.426	0.782	10	12,16,14	0.458	0.778
Mean values		**0.451**	**0.735**			**0.514**	**0.775**

Third year				Fourth year			
k	Variables	Accuracy	Cost	k	Variables	Accuracy	Cost
1	12,14,16	0.522	0.707	1	12,14,16	0.551	0.662
2	12,11,14	0.500	0.797	2	12,14,6	0.564	0.759
3	12,11,14	0.544	0.623	3	16,12,14	0.493	0.753
4	12,16,14	0.533	0.752	4	12,16,14	0.525	0.701
5	12,14,7	0.533	0.685	5	12,16,14	0.558	0.636
6	12,14,6	0.561	0.775	6	12,16,14	0.538	0.655
7	12,11,14	0.556	0.646	7	12,16,14	0.545	0.668
8	12,14,7	0.522	0.752	8	12,14,4	0.506	0.766
9	12,14,4	0.511	0.707	9	12,16,14	0.493	0.798
10	12,14,4	0.533	0.634	10	12,14,6	0.538	0.655
Mean values		**0.532**	**0.708**			**0.533**	**0.706**

accuracy and cost for the *CS-SNB-Cost* model by each fold of the cross-validation process. These values were achieved using the cost matrix $C(0, n^2)$ for all prediction years. Analyzing Table 6.5, it is found that the models always include the *impact factor* (variable 12). The models also usually include other variables like the h_c*-index* (variable 11), the *cited half-life* (variable 14) and the *article influence* (variable 16). It is also noted that fewer models include the *g-index* (variable 4), the *a-index* (variable 6) and the *m-index* (variable 7). Note that first-year models always had two variables, whereas second-, third-, and fourth-year models almost always had three variables. Finally, it is found that the feature variables of the model that was most often induced included *impact factor*, *cited half-life* and *article influence*. Not all the models included these variables in all situations. This depends on the cost matrix and prediction year. So, other models were formed by different variables, although the *impact factor*, the *cited half-life* and the *article influence* were also present.

6.3.4 Exploiting the proposed models

The increase of the *h-index* value of a Neurosciences journal in the first year was predicted as an example. Table 6.6 shows the parameters that define the model which are learned using the cost matrix $C(0, n)$. All features are described by means of the mean (μ) and the

Table 6.6: Parameters that define a specific cost-sensitive selective naive Bayes classifier for first-year models. Feature variables belonging to this classifier are: impact factor, cited half-life and article influence.

	impact factor	cited half-life	article influence
Δh=low	μ=0.977 σ=0.906	μ=5.573 σ=3.328	μ=0.318 σ=0.354
Δh=medium-low	μ=1.781 σ=1.021	μ=6.335 σ=2.394	μ=0.608 σ=0.414
Δh=medium-high	μ=2.728 σ=1.724	μ=5.985 σ=2.205	μ=0.950 σ=0.830
Δh=high	μ=5.774 σ=4.802	μ=5.643 σ=2.060	μ=2.660 σ=3.202

Table 6.7: Results predicted by cost-sensitive selective naive Bayes classifiers (*CS-SNB-Accuracy* and *CS-SNB-Cost*) for different years

	First year	Second year	Third year	Fourth year
CS-SNB-Accuracy				
low	0.076	0.212	0.126	0.101
medium-low	0.391	**0.387**	**0.454**	**0.477**
medium-high	**0.473**	0.317	0.338	0.324
high	0.060	0.084	0.082	0.098
CS-SNB-Cost				
low	0.070	0.159	0.126	0.101
medium-low	0.291	**0.410**	**0.454**	**0.477**
medium-high	**0.504**	0.330	0.338	0.324
high	0.135	0.101	0.082	0.098

standard deviation (σ).

Given a journal ($\mathbf{x}$) with the following values: *impact factor*=2.582, *cited half-life*=5.6, and *article influence*=0.852, the Δh values can be predicted using the formulation of cost-sensitive selective naive Bayes (Algorithm 1 and Algorithm 2) and the parameters listed in Table 6.6.

After propagating the above evidence, the results predicted by *CS-SNB-Accuracy* model were $p(\Delta h$=*low* $|\boldsymbol{x})$=0.076, $p(\Delta h$=*medium-low* $|\boldsymbol{x})$=0.391, $p(\Delta h$=*medium-high* $|\boldsymbol{x})$=0.473 and $p(\Delta h$=*high* $|\boldsymbol{x})$=0.060. Similarly, the *CS-SNB-Cost* model predicted $p(\Delta h$=*low* $|\boldsymbol{x})$=0.070, $p(\Delta h$=*medium-low* $|\boldsymbol{x})$=0.291, $p(\Delta h$=*medium-high* $|\boldsymbol{x})$=0.504 and $p(\Delta h$=*high* $|\boldsymbol{x})$=0.135. According to both approaches the *h-index* of the above journal is likely to increase by three or four units (medium-high) in the next year.

Table 6.7 shows other prediction years using the above conditions. It is observed that both models predicted the same class for all years. Note that accuracy is different for first- and second-year models, but the same for third- and fourth-year models are equals. This is because the induced first- and second-year models were different. Results shows that the increase of *h-index* for the above journal ($\mathbf{x}$) will be medium-high (Δh=[3-4]) in the first year, and medium-low in the second (Δh=[5-6]), third (Δh=[7-9]) and fourth (Δh=[9-12]) years.

6.4 Discussion and conclusions

Machine learning community is not only interested in maximizing classification accuracy, but also in minimizing the expected total cost associated with misclassifications. Some ideas, like the cost-sensitive learning approach, are proposed to face this problem. In this chapter, two greedy wrapper forward cost-sensitive selective naive Bayes approaches are proposed. Both approaches include the misclassification costs in the learning process, readjusting the probability thresholds of each class to select the class with the minimum-expected cost. In this context, the search process of the first approach (*CS-SNB-Accuracy*) includes variables that improve classification accuracy of the model, whereas the search process of the second approach (*CS-SNB-Cost*) includes variables that reduce the distances between the actual and the predicted classes.

These proposed algorithms have been tested on the bibliometric indices prediction area. Considering the popularity of the well-know *h-index*, several prediction models, based on the cost-sensitive approach and the feature subset selection, are learned to forecast the annual increase of the *h-index* for Neurosciences journals in a four-year time horizon. Models capable of predicting the *h-index* that a scientific journal is likely to have in coming years can be a useful tool for the scientific community.

Results show that our approaches, specially the *CS-SNB-Accuracy*, achieved higher accuracy values than the analyzed cost-sensitive classifiers and Bayesian classifiers. Furthermore, we also noted that the *CS-SNB-Cost* always achieved a lower average cost than all analyzed cost-sensitive and cost-insensitive classifiers. These cost-sensitive selective naive Bayes approaches outperform the original selective naive Bayes in terms of accuracy and average cost, so the cost-sensitive learning approach could be used in different probabilistic classification approaches. In the future, new cost-sensitive Bayesian classifiers like the selective tree augmented naive Bayes could include other journal-based features, e.g., 5-year impact factor, percentage of documents published by a single author, percentage of documents published in international collaboration and so on.

Chapter 7

Discovering relationships among indices using Bayesian networks

7.1 Introduction

Although bibliometric indices are usually used to evaluate the impact of a researcher's work, they are also used as a tool for journal evaluation [171]. The first bibliometric index to be calculated for journal assessment was the *impact factor* [172]. More recently, Braun [60] suggested that the well-known *h-index* could be usefully applied to evaluate the scientific impact of journals. Other indices like *eigenfactor*, *article influence* and *Scimago's journal rank index*, among others, have also been developed for the same purpose.

In view of the vast number of bibliometric indices, it is necessary to analyze how they relate to each other. The degree of correlation among journal citation indices has been investigated in the past. Some studies have examined correlations between a list of journal citation indices using the Pearson ρ coefficient. Bollen [48] found statistically significant correlations between 39 measures of scholarly impact. Franceschet [154] made a thorough comparison of the rankings provided by *Eigenfactor* and *5-year impact factor*. The author found that, although the two bibliometric measures are generally statistically correlated, they also significantly diverge in some cases. Furthermore, Leydesdorff [287] showed high correlations between indices, specially the *5year impact factor* and *article influence* (ρ=0.956). Similarly, other studies have also examined the degree of correlation between some typical journal citation indices, tested using Spearman's ρ. For example, Bollen *et al.* [47] compared journal *PageRank* with *2-year impact factor*. They found a moderate correlation (ρ=0.63) between rankings for computer science journals. Chen *et al.* [86] and Ma *et al.* [301] found a Spearman correlation of 0.91 and 0.98, respectively, between the article rankings provided by PageRank and total *number of citations*. Davis [114] compared the rankings according to *eigenfactor* and *2-year impact factor*. The author found a significant correlation between the two measures (ρ=0.84), and an even higher association (ρ=0.95) between *eigenfactor* and the total *number of citations*. Also, Franceschet [155] showed strong correlations between the

2year impact factor and the *5year impact factor* (ρ=0.96). Finally, Saad [389] noticed that the *2year impact factor* was more correlated with *article influence* than with *eigenfactor*.

To date there have not been many publications analyzing citations in computer science and artificial intelligence. In this context, Serenko [409] analyzed journals in the field of artificial intelligence, and calculated some bibliometric index' values (*h-index*, *g-index* and *hc-index*) which correlated almost perfectly with each other (ranging from 0.97 to 0.99). It should be mentioned that most bibliometric indices are obviously correlated since they are all derived from the number of documents and citations, and these are highly correlated.

The interest and originality of this analysis is that it introduces a new Bayesian network-based approach for analyzing the conditional (in)dependencies between journal citation indices. Some Bayesian networks (yearly and global models) are learned to discover the relationships between 14 journal citation indices. These models are built using journal publication and citation data (all the journals in the JCR Computer Science and Artificial Intelligence category) during the period 2000-2009, inclusive. Yearly and global models are developed to analyze index relationships within a one-year publication and citation window. Finally, it is analyzed how some indices influence others in probabilistic terms. Also, the network is able to perform all kinds of probabilistic reasoning, computing, say, the probability of a journal obtaining certain fixed index values given other known values.

The main advantage of this analysis over earlier studies, which analyze only bivariate correlations between indices, is that it calculates the joint probability distribution over all analyzed indices, discovering probabilistic conditional (in)dependencies among triplets of indices. Journal citation indices have never been analyzed like this before. Using the proposed models, computer science and artificial intelligence journal editorial boards could answer some of the questions related to their journal citation indices, like, for example, what would happen to our journal's *impact factor* if papers, which are published by our journal, received more *citations*?, what would happen to our journal's *h-index* if our journal published a lot of new *documents*?, what would happen to our journal's *g-index* if our most cited papers, were the only ones to receive new citations? or what would happen to our journal's *immediacy-index* if our journal accepted more *documents*, but received no new *citations*?

Obviously, this is a general-purpose methodology and can be used for other research areas, and, the editorial boards of any journals could find answers to the above questions. This work appears in the published paper [225].

Chapter outline

The remainder of the chapter is organized as follows. Section 7.2 presents the dataset used, the Bayesian networks learned, the discovered probabilistic conditional (in)dependencies among the analyzed indices and examples of probabilistic reasoning. Finally, Section 7.3 contains some conclusions emphasizing the original contribution of the chapter and future research on the topic.

7.2 Analyzing conditional (in)dependencies among indices

Bayesian network models are selected to discover conditional independencies among bibliometric indices. In particular, each node in the network represents a specific index, while the arcs between indices represent the conditional (in)dependencies among these indices. By learning these Bayesian networks from data, the aim is to discover probabilistic conditional (in)dependencies among the set of bibliometric indices. The indices analyzed in this study are: *documents*, *citations*, *h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *w-index*, h_r*-index*, *impact factor* and *immediacy-index*. All these indices were obtained from the information provided by the Web of Science. Despite most of the above indices were originally developed to evaluate the quality of a researcher's work, they have been adapted to assess journals.

7.2.1 Dataset compilation

Web of Science platform is selected as source to download publication and citation data. Firstly, journal data from the JCR's Computer Science and Artificial Intelligence category is collected. There are 94 journals in this category of the 2008 Journal Citation Reports Science Edition. In view of the objective of this analysis, only the 70 journals (Table 7.1) that published papers from January 1, 2000 to December 31, 2009 were took into account. The next step was to obtain the publication list and citation data for these journals. Finally, the last step was to use all the above information to calculate some scientific impact indices (*documents*, *citations*, the *h-index*, the *g-index*, the *hg-index*, the *a-index*, the *m-index*, the q^2*-index*, the *r-index*, the *e-index*, the *w-index*, the *rational h-index*, the *impact factor* and the *immediacy-index*) associated with the selected journals. These index values have been calculated yearly for each of the 70 journals in the ten-year period from 2000 to 2009.

7.2.2 Data distribution

To illustrate some of the calculated index values, Table 7.2 lists some of the journals ranked top according to six selected indices: *documents*, *citations*, the *h-index*, the *g-index*, the *impact factor* and the *immediacy-index*. Table 7.2 shows some rankings obtained using data for a one-year publication window, specifically for 2009. These rankings reveal that some journals are always positioned near to the top. Taking *IEEE Transactions on Pattern Analysis and Machine Intelligence* as an example, it is found that this journal published 188 papers in 2009, which received 54 citations in the same year. Furthermore, its *h-index*' value and *g-index*' value were 3 and 4, respectively. Finally, it had an *impact factor* value of 5.960 and an *immediacy-index* value of 0.669. Remember that all these values were obtained using a 2009-year publication window.

Table 7.3 illustrates the range of index values in each analyzed year. Taking the *documents* value for the year 2000 as an example, it is found that the minimum number of documents published by a specific journal in the year 2000 was 10 and the maximum was 219. Analyzing this table, it is found that index values are higher for recent than older years. Specifically,

Table 7.1: List of journals in JCR Computer Science and Artificial Intelligence category that have papers every year throughout the 2000-2009 period.

Journals
Adaptive Behavior
AI Communications
AI Edam-Artificial Intelligence for Engineering Design Analysis and Manufacturing
AI Magazine
Annals of Mathematics and Artificial Intelligence
Applied Artificial Intelligence
Applied Intelligence
Artificial Intelligence
Artificial Intelligence in Medicine
Artificial Intelligence Review
Artificial Life
Autonomous Agents and Multi-Agent Systems
Autonomous Robots
Chemometrics and Intelligent Laboratory Systems
Computational Intelligence
Computer Speech and Language
Computer Vision and Image Understanding
Connection Science
Data and Knowledge Engineering
Data Mining and Knowledge Discovery
Decision Support Systems
Engineering Applications of Artificial Intelligence
Engineering Intelligent Systems for Electrical Engineering and Communications
Expert Systems
Expert Systems with Applications
IEEE Transactions on Evolutionary Computation
IEEE Transactions on Fuzzy Systems
IEEE Transactions on Image Processing
IEEE Transactions on Knowledge and Data Engineering
IEEE Transactions on Neural Networks
IEEE Transactions on Pattern Analysis and Machine Intelligence
IEEE Transactions on Systems, Man and Cybernetics Part B-Cybernetics
IEEE Transactions on Systems, Man and Cybernetics Part C-Applications and Reviews
Image and Vision Computing
International Journal of Approximate Reasoning
International Journal of Computer Vision
International Journal of Intelligent Systems
International Journal of Patter Recognition and Artificial Intelligence
International Journal of Software Engineering and Knowledge Engineering
International Journal of Uncertainty Fuzziness and Knowledge-Based Systems
Integrated Computer-Aided Engineering
Intelligent Automation and Soft Computing
Journal of Artificial Intelligent Research
Journal of Automated Reasoning
Journal of Chemometrics
Journal of Computer and Systems Sciences International
Journal of Experimental and Theoretical Artificial Intelligence
Journal of Heuristics
Journal of Intelligent and Fuzzy Systems
Journal of Intelligent Information Systems
Journal of Intelligent Manufacturing
Journal of Intelligent and Robotic Systems
Journal of Mathematical Imaging and Vision
Knowledge Engineering Review
Knowledge-Based Systems
Machine Learning
Machine Vision and Applications
Mechatronics
Medical Image Analysis
Minds and Machines
Network-Computation in Neural Systems
Neural Computation
Neural Computing and Applications
Neural Networks
Neural Processing Letters
Neurocomputing
Pattern Analysis and Applications
Pattern Recognition
Pattern Recognition Letters
Robotics and Autonomous Systems

Table 7.2: Top five positions of the journal rankings using a 2009-year publication and citation window according to six bibliometric indices

Position	Journal	*Documents*
1	Expert Systems with Applications	1399
2	Neurocomputing	352
3	Pattern Recognition	312
4	IEEE Transactions on Image Processing	217
5	IEEE Transactions on Pattern Analysis and Machine Intelligence	188

Position	Journal	*Citations*
1	Expert Systems with Applications	402
2	Neurocomputing	68
3	IEEE Transactions on Pattern Analysis and Machine Intelligence	54
4	Neural Networks	49
5	Pattern Recognition	45

Position	Journal	*h-index*
1	Expert Systems with Applications	5
2	Neurocomputing	4
3	Neural Networks	4
4	IEEE Transactions on Pattern Analysis and Machine Intelligence	3
5	Image and Vision Computing	2

Position	Journal	*g-index*
1	Expert Systems with Applications	6
2	Neurocomputing	5
3	Neural Networks	4
4	IEEE Transactions on Pattern Analysis and Machine Intelligence	4
5	IEEE Transactions on Knowledge and Data Engineering	3

Position	Journal	*impact factor*
1	IEEE Transactions on Pattern Analysis and Machine Intelligence	5.960
2	International Journal of Computer Vision	5.358
3	IEEE Transactions on Evolutionary Computation	3.736
4	IEEE Transactions on Neural Networks	3.726
5	IEEE Transactions on Fuzzy Systems	3.624

Position	Journal	*immediacy-index*
1	Computational Intelligence	1.091
2	IEEE Transactions on Pattern Analysis and Machine Intelligence	0.669
3	Artificial Intelligence	0.667
4	International Journal of Computer Vision	0.659
5	Journal of Automated Reasoning	0.600

most of the highest values for each index are obtained between 2007 and 2009. Although it is observed that the values of all indices tend to increase, there is no index whose value increases year by year.

Note that the highest values differ greatly depending on the selected index and year. On the one hand, the values of indices like *documents* or *citations* have undergone a more significant increase than other indices over the analyzed years. The number of *documents* and *citations* incremented sharply in the time period. In fact, they are six and nine times greater, respectively, in the last year than in the first year. On the other hand, the values of other indices (*h-index*, *g-index*, *m-index*, *w-index*, *rational h-index*, *impact factor*) have not increased as significantly as *documents* and *citations*. In the most recent years, they are approximately two times greater than in the early years. For example, the *g-index* has a value of 3 in the year 2000 and a value of 6 in 2009. In the same way, the *impact factor* has a

Table 7.3: Range of index values in each analyzed year. Numbers in boldface represent the maximum value for each index in the 2000-2009 period.

Index	2000		2001		2002		2003		2004	
	min	max	min	max	min	max	min	max	min	max
documents	10	219	12	257	8	308	15	297	12	295
citations	0	45	0	47	0	39	0	78	0	48
h-index	0	3	0	2	0	3	0	4	0	3
g-index	0	3	0	4	0	3	0	4	0	4
hg-index	0.0	3.0	0.0	2.8	0.0	3.0	0.0	4.0	0.0	3.5
a-index	0.0	5.0	0.0	6.5	0.0	4.0	0.0	5.0	0.0	6.0
m-index	0	3	0	4	0	3	0	4	0	3
q^2*-index*	0.0	3.0	0.0	2.8	0.0	3.0	0.0	4,0	0.0	3,0
r-index	0.0	3.5	0.0	3.6	0.0	3.5	0.0	4.5	0.0	3.9
e-index	0.0	2.0	0.0	3.0	0.0	1.7	0.0	2.4	0.0	2.6
w-index	0	4	0	4	0	5	0	6	0	5
rational h-index	0.0	3.6	0.0	2.8	0.0	3.7	0.0	4.8	0.0	3.6
impact factor	0.0	2.8	0.0	2.7	0.0	2.7	0.0	2.9	0.0	4.4
immediacy-index	0.0	0.7	0.0	1.0	0.0	0.7	0.0	0.9	0.0	0.7

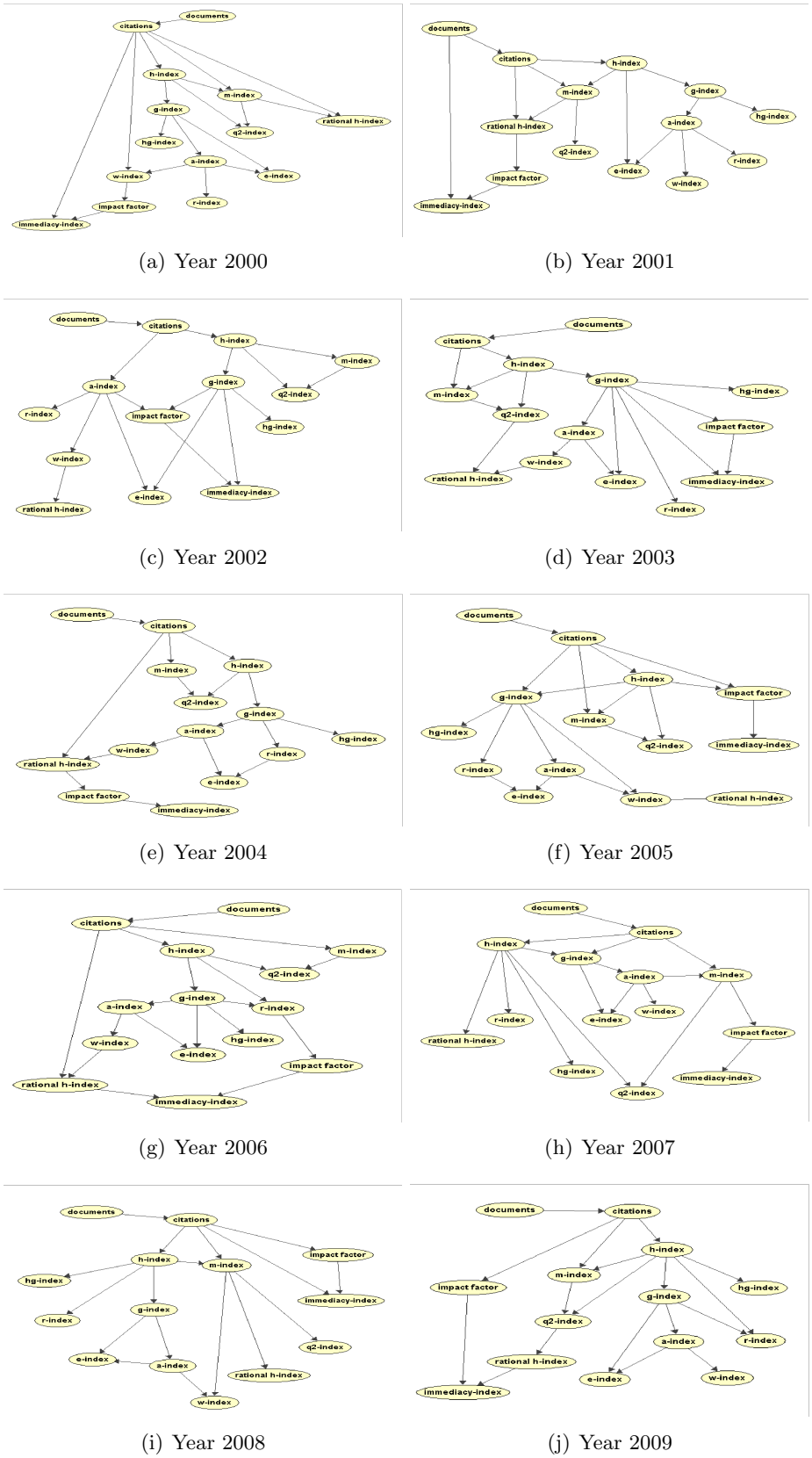
Index	2005		2006		2007		2008		2009	
	min	max	min	max	min	max	min	max	min	max
documents	15	255	18	335	16	342	3	523	3	**1399**
citations	0	86	0	81	0	73	0	164	0	**402**
h-index	0	4	0	4	0	3	0	4	0	**5**
g-index	0	5	0	4	0	5	0	5	0	**6**
hg-index	0.0	4.5	0.0	4.0	0.0	3.9	0.0	4.5	0.0	**5.0**
a-index	0.0	6.0	0.0	5.0	0.0	**6.7**	0.0	6.5	0.0	6.6
m-index	0	5	0	4	0	4	0	5	0	**6**
q^2*-index*	0.0	4.5	0.0	4.0	0.0	3.5	0.0	4.5	0.0	**5.5**
r-index	0.0	4.9	0.0	4.2	0.0	4.5	0.0	4.8	0.0	**5.7**
e-index	0.0	2.8	0.0	2.4	0.0	**3.3**	0.0	3.0	0.0	2.8
w-index	0	7	0	5	0	6	0	7	0	**8**
rational h-index	0.0	4.9	0.0	4.4	0.0	3.9	0.0	4.9	0.0	**5.8**
impact factor	0.1	4.3	0.0	3.8	0.1	**6.1**	0.0	3.6	0.0	6.0
immediacy-index	0.0	1.0	0.0	1.0	0.0	**1.2**	0.0	0.8	0.0	1.1

value of 2.8 in the year 2000 and a value of 6.0 in 2009. Finally, the values of the other indices (*hg-index*, *a-index*, q^2*-index*, *r-index*, *e-index*, *immediacy-index*) have also increased within the time period, but to a lesser extent than the other indices. For example, the *a-index* has a value of 5.0 in the year 2000 and a value of 6.6 in 2009. Similarly, the *immediacy-index* has a value of 0.7 in the year 2000 and a value of 1.1 in 2009.

7.2.3 Bayesian network models

Yearly Bayesian network models Ten yearly Bayesian networks models were learned to analyze the relationships among indices within the same one-year publication window. For this reason, each yearly model is associated with one of the ten analyzed years and with one of the ten datasets. Each dataset contains 70 cases (journals), each one with its 14 index values.

Before running the K2 algorithm to learn a Bayesian network from each dataset, some K2's requirements are established. Since K2 needs the variables to be ordered, the first decision was to specify an order. Taking into account the index definitions, indices that could be parents of the other indices were placed first. The established order was: *documents*,



(a) Year 2000

(b) Year 2001

(c) Year 2002

(d) Year 2003

(e) Year 2004

(f) Year 2005

(g) Year 2006

(h) Year 2007

(i) Year 2008

(j) Year 2009

Figure 7.1: Bayesian network structures learned for each analyzed year

citations, *h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *w-index*, *rational h-index*, *impact factor* and *immediacy-index*. High values of marginal likelihood were obtained using this order. The second requirement was to assign a value to the maximum number of parents. This was set at two due to the dataset characteristics. There was a third requirement: index values had to be discretized into intervals. Due to the number of dataset cases, values were discretized into three intervals with equal frequency. In this way, the index values were assigned to one of the three possible values (*low*, *medium* and *high*).

According to the network structures, Figure 7.2.3 presents the ten yearly Bayesian networks. There are a lot of coincident arcs in the yearly Bayesian networks as shown in Table 7.4. Taking the value of the first-row and second-column as an example, the value displayed is 14. This value indicates that the Bayesian network for the year 2000 and the Bayesian network for the year 2001 have 14 identical arcs. In other words, the Bayesian network of the year 2000 has 19 arcs, and 14 of these arcs are also represented in the Bayesian network of the year 2001. Examining Table 7.4, it is found that the relationships between indices are very similar in each year of the analyzed period.

Analyzing the networks in Figure 7.2.3, it is found that there are some specific arcs that are represented in most of the networks. For example, the arcs *h-index*→*g-index*, *documents*→*citations*, *citations*→*h-index*, *a-index*→*e-index*, among others, always appear in the 10 yearly Bayesian networks. The number of times that an arc is shown in the proposed Bayesian networks is reported in Table 7.5. Taking the value of the *citations*→*m-index* arc as an example, the value displayed is 9. This value means that the relationship between *citations* and *m-index* is present in 9 out of our 10 yearly Bayesian networks.

With the intention of representing the main relationships between indices in a single year-independent Bayesian network, an aggregated Bayesian network was built using only those arcs that had appeared at least three times. After applying the above filter, it is obtained the aggregated Bayesian network shown in Figure 7.2. The values above the arcs represent the number of times that the arc appeared in the ten yearly Bayesian networks.

Examining the index definitions, it is observed that some of them can be defined according to the values of other indices. For example, on the one hand, the *hg-index* can be expressed in terms of *h-* and *g-index*'s values ($hg\text{-}index = \sqrt{h \cdot g}$) and, on the other hand, the q^2*-index* can be defined according to *h-* and *m-index*'s values ($q^2\text{-}index = \sqrt{h \cdot m}$). Although

Table 7.4: Number of coincident arcs in the 10 different networks

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
2000 (19 arcs)	-	14	12	13	12	12	13	11	13	12
2001 (18 arcs)	-	-	10	11	12	11	11	9	11	10
2002 (17 arcs)	-	-	-	14	10	11	11	9	9	10
2003 (18 arcs)	-	-	-	-	13	14	14	11	11	14
2004 (16 arcs)	-	-	-	-	-	14	14	10	9	11
2005 (19 arcs)	-	-	-	-	-	-	13	11	11	13
2006 (18 arcs)	-	-	-	-	-	-	-	12	11	14
2007 (17 arcs)	-	-	-	-	-	-	-	-	12	13
2008 (17 arcs)	-	-	-	-	-	-	-	-	-	14
2009 (18 arcs)	-	-	-	-	-	-	-	-	-	-

Table 7.5: Number of times that arcs appear in our 10 yearly Bayesian networks

	documents	*citations*	*h-index*	*g-index*	*hg-index*	*a-index*	*m-index*	*q^2-index*	*r-index*	*e-index*	*w-index*	*rational h-index*	*impact factor*	*immediacy-index*
documents	-	10	0	0	0	0	0	0	0	0	0	0	0	0
citations	-	-	10	2	0	1	9	0	0	0	1	4	3	2
h-index	-	-	-	10	3	0	7	8	4	1	0	1	1	0
g-index	-	-	-	-	7	9	0	0	5	7	1	0	2	2
hg-index	-	-	-	-	-	0	0	0	0	0	0	0	0	0
a-index	-	-	-	-	-	-	1	0	3	10	10	0	1	0
m-index	-	-	-	-	-	-	-	10	0	0	1	3	1	0
q^2-index	-	-	-	-	-	-	-	-	0	0	0	2	1	0
r-index	-	-	-	-	-	-	-	-	-	2	0	0	1	0
e-index	-	-	-	-	-	-	-	-	-	-	0	0	0	0
w-index	-	-	-	-	-	-	-	-	-	-	-	5	1	0
rational h-index	-	-	-	-	-	-	-	-	-	-	-	-	2	2
impact factor	-	-	-	-	-	-	-	-	-	-	-	-	-	10
immediacy-index	-	-	-	-	-	-	-	-	-	-	-	-	-	-

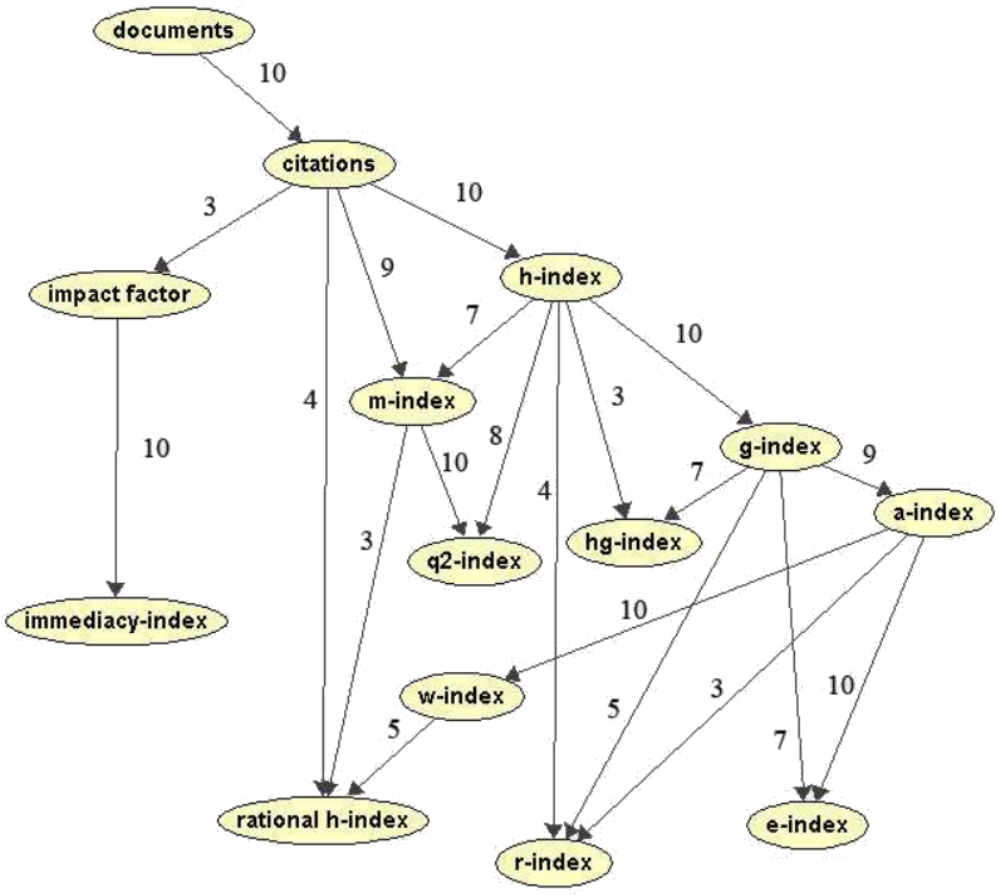


Figure 7.2: Aggregated Bayesian network structure

the deterministic aspect between index definitions is reduced after discretizing the index values into three intervals (*low*, *medium* and *high*), the proposed aggregated Bayesian network discovers such dependencies between indices. Looking at Figure 7.2, it is found that the dependencies that can be defined according to the values of other indices are represented in the aggregated Bayesian network. For example, the *hg-index* definition is represented in the network since the *h-index* and the *g-index* are parents of the *hg-index* in the network. Similarly, the q^2*-index* as a function of the *h-index* and the *m-index* is also represented in the network.

Some dependencies, like *documents*→*citations* and *h-index*→*g-index*, among others, are not derived from the index definition, but were expected because many works showed their correlations [101, 396]. Other dependencies, e.g., *citations*→*h-index* and *citations*→*impact factor* are also represented in the aggregated Bayesian network. Note that although the *h-index* and the *impact factor* cannot be defined in terms of *citations* values only, they do exhibit a high value correlation coefficient [57]. Other dependencies, e.g., the arc between *a-index* and *e-index*, is an example of a dependency that was not initially expected. Remember that the *a-index* represents the average number of citations received by the articles included in the *h-core*, whereas the *e-index* represents the excess citations received by the articles in the *h-core*. Thus, both refer to citations of articles in the *h-core*. More examples of such dependencies are: *m-index*→*rational h-index*, *a-index*→*w-index*, and *w-index*→*rational h-index* among others.

In order to discover conditional independencies among the analyzed indices, Markov properties were used as the criteria for this purpose. The local Markov property states that any node in any Bayesian network is conditionally independent of its *non-descendants* given its *parents*, whereas the global Markov property states that any node is conditionally independent of any other node given its Markov blanket (MB) which includes its parents, its children, and its children's parents. Table 7.6 illustrates such relationships. Although this table shows a specific list of relationships between indices, new relationships can be derived using other conditional independencies properties [81]:

- Symmetry: $I(X, Y \mid Z) \Leftrightarrow I(Y, X \mid Z)$
- Decomposition: $I(X, (Y \cup W) \mid Z) \Rightarrow I(X, Y \mid Z)$ and $I(X, W \mid Z)$
- Strong joint: $I(X, Y \mid Z) \Rightarrow I(X, Y \mid (Z \cup W))$

Taking the q^2*-index* as an example, it is found in Table 7.6 that, given the *h-index* and the *m-index* together, the q^2*-index* is independent of most of the indices (*documents*, *citations*, the *hg-index*, the *a-index*, the *r-index*, the *e-index*, the *w-index*, the *rational h-index*, the *impact factor* and the *immediacy-index*). On the other hand, it is observed that the *g-index* is independent of the q^2*-index* given the *h-index*. Taking into account the above independency relationships and the conditional independency properties, we state that given the *h-index* and the *m-index* together, the q^2*-index* is independent of any of the other indices. This means that when we know the *h-index* and the *m-index* values, knowledge of the *e-index*,

Table 7.6: Conditional independencies among indices, derived using Markov properties in the aggregated Bayesian network

Index	is conditionally independent of	given
documents	*h-index*, *g-index*,*hg-index*, *a-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *w-index*, *rational h-index*, *impact factor*, *immediacy-index*	*citations*
citations	*hg-index*, *a-index*, *r-index*, *e-index*, *immediacy-index*	*documents*, *h-index*, *m-index*, *w-index*, *rational h-index*, *impact factor*
h-index	*immediacy-index*	*citations*
h-index	*rational h-index*	*citations*, *g-index*, *hg-index*, *a-index*, *m-index*, q^2*-index*, *r-index*
g-index	*documents*, *citations*, *m-index*, q^2*-index*, *impact factor*, *immediacy-index*	*h-index*
g-index	*rational h-index*	*h-index*, *hg-index*, *e-index*, *a-index*, *r-index*
hg-index	*documents*, *citations*, *m-index*, q^2*-index*, *r-index*, *e-index*, *w-index*, *rational h-index*, *impact factor*, *immediacy-index*	*h-index*, *g-index*
a-index	*documents*, *citations*, *h-index*, *hg-index*, *m-index*, q^2*-index*, *impact factor*, *immediacy-index*	*g-index*
a-index	*rational h-index*	*h-index*, *g-index*, *r-index*, *e-index*, *w-index*
m-index	*hg-index*, *a-index*, *r-index*, *e-index*, *w-index*, *immediacy-index*	*citations*, *h-index*
q^2*-index*	*documents*, *citations*, *hg-index*, *a-index*, *r-index*, *e-index*, *w-index*, *rational h-index*, *impact factor*, *immediacy-index*	*h-index*, *m-index*
r-index	*documents*, *citations*, *m-index*, q^2*-index*, *rational h-index*, *impact factor*, *immediacy-index*	*h-index*, *g-index*, *a-index*
e-index	*documents*, *citations*, *h-index*, *hg-index*, *m-index*, q^2*-index*, *r-index*, *rational h-index*, *impact factor*, *immediacy-index*	*g-index*, *a-index*
w-index	*documents*, *citations*, *h-index*, *g-index*, *hg-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *impact factor*, *immediacy-index*	*a-index*
rational h-index	*h-index*, *g-index*, *hg-index*, *a-index*, q^2*-index*, *r-index*, *e-index*, *immediacy-index*	*citations*, *m-index*, *w-index*
impact factor	*h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *w-index*, *rational h-index*	*citations*
immediacy-index	*documents*, *citations*, *h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *w-index*, *rational h-index*	*impact factor*

for example, provides no information on the occurrence of the q^2*-index*. Similarly, note that given the *h-index* and the *g-index* together, the *hg-index* is independent of any of the other indices.

The aggregated Bayesian network is able to encode the conditional independencies that are derived from the index definition, but also discovers other new conditional independencies that are not strictly derived from definitions. On the one hand, it is expected the *m-index*, which is the median number of citations received by papers in the *h-core*, to be conditionally independent of some variables given *citations* and *h-index*. Table 7.6 shows that given *citations* and *h-index*, the *m-index* is conditionally independent of *hg-index*, *a-index*, *r-index*, *e-index*, *w-index* and *immediacy-index*. On the other hand, some other conditional independencies were not so obvious. For example, the *immediacy-index*, which is defined by means of *documents*' values and *citations*' values, is independent of *documents* and *citations* given the *impact factor*. The presence of *impact factors* values has a significant influence on *immediacy-index* values. This means that when we know the *impact factor*, knowledge of *documents* and *citations* does not provide any information on the occurrence of the *immediacy-index*. Remember that the conditional independencies between indices encoded in the proposed Bayesian networks do not represent a causality relationship, but refer to a probability relationship between indices.

Global Bayesian network model The global model's objective is to analyze the relationships between indices within a one-year target window. It has the same as the yearly models, but, unlike them, this global model is built using a different dataset. Previous datasets are merged into a single dataset to build the dataset for the global model. In this way, the global model dataset contains 700 cases. Each case in the dataset was referred to index values calculated within a one-year target window, and, consequently, they all cases can be easily merged into a single dataset.

The K2 scoring metric was also used to learn the global Bayesian network. Although the same variables order is established, some decisions about K2's requirements are modified. As there are more cases, it can be increased the maximum number of parents for any node up to 3 and the number of intervals for the index discrete domain to 4 (*low*, *medium-low*, *medium-high* and *high*).

Figure 7.3 shows our global Bayesian network structure. Comparing the arcs in Figure 7.2 (aggregated Bayesian network structure) and Figure 7.3 (global Bayesian network structure), it is found that the network structures are similar. There are many arcs that appear in both networks, although this new model includes some specific arcs not represented before: *citations*→*a-index*, *h-index*→*a-index*, *h-index*→*e-index*, *h-index*→*rational h-index*, among others. For this reason, it is expected that the global model more accurately represents the index definition than the aggregated model. These new dependencies are explained in the following.

Some centrality measures are examined in order to analyze some of the index characteristics in the proposed global Bayesian network. Centrality degree is defined as the number of arcs incident upon an index. Degree is often interpreted in terms of the opportunity for influencing any other index. Two separate measures of centrality degree (indegree and outdegree) are defined. A node's indegree is the number of arcs directed to the node, and

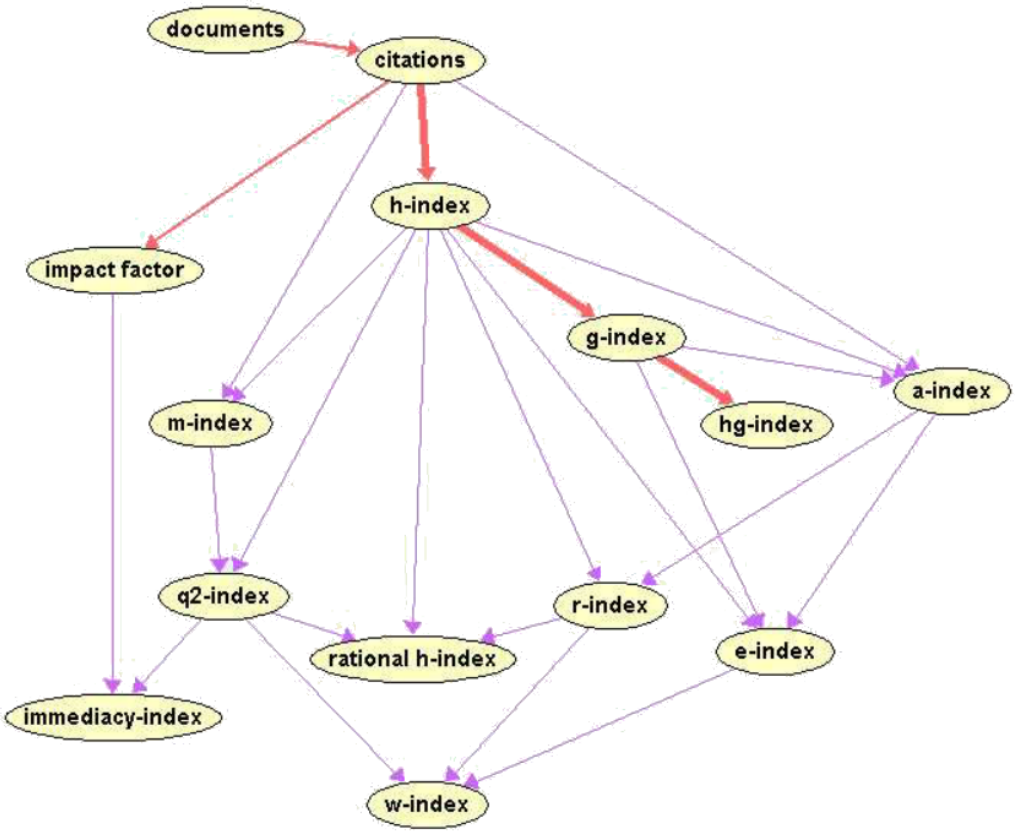


Figure 7.3: Global Bayesian network structure

outdegree is the number of arcs that the node directs to others. Therefore, indegree is the number of parents, whereas outdegree is the number of children. The centrality degree (CD) values are: CD(*documents*)=1, CD(*citations*)=5, CD(*h-index*)=8, CD(*g-index*)=4, CD(*hg-index*)=1, CD(*a-index*)=5, CD(*m-index*)=3, CD(q^2*-index*)=5, CD(*r-index*)=4, CD(*e-index*)=4, CD(*w-index*)=3, CD(*rational h-index*)=3, CD(*impact factor*)=2 and CD(*immediacy-index*)=2. Examining the above values, it is observed that the *h-index* has a lot of influence on other indices. It has the highest degree centrality value (1+7=8). On the other hand, indices like *documents* or *hg-index*, which have a centrality degree of 1, do not influence the other indices so much.

Table 7.7 displays the range of index values and the values assigned to each interval after the discretization to illustrate the global model dataset. Note that the interval width is not the same for the four categories since the dataset has been discretized in intervals of equal frequency. A journal publishes a number of between 3 and 1399 *documents* per year, and these journals can then be categorized according to the index values. For example, a journal that has published 43 documents per year is placed in the *medium-low* category.

After examining the dependency relationships between indices shown in Figure 7.3, it is found that the aggregated model identified some, but not all, the index dependencies. Note

Table 7.7: Range of index values related to each interval of the global model dataset

Indices	Values' range	*low*	*medium-low*	*medium-high*	*high*
documents	[3, 1399]	[3, 29)	[29, 47)	[47, 79)	[79, 1399]
citations	[0, 402]	[0, 1)	[1, 3)	[3, 12)	[12, 402]
h-index	[0, 5]	[0, 1)	[1, 2)	[2, 3)	[3, 5]
g-index	[0, 6]	[0, 1)	[1, 2)	[2, 3)	[3, 6]
hg-index	[0.0, 5.5]	[0.0, 1.0)	[1.0, 1.4)	[1.4, 2.4)	[2.4, 5.5]
a-index	[0.0, 6.7]	[0.0, 1.0)	[1.0, 2.0)	[2.0, 3.0)	[3.0, 6.7]
m-index	[0, 6]	[0.0, 1.0)	[1.0, 2.0)	[2.0, 3.0)	[3.0, 6]
q^2*-index*	[0.0, 5.5]	[0.0, 1.0)	[1.0, 1.4)	[1.4, 2.0)	[2.0, 5.5]
r-index	[0.0, 5.7]	[0.0, 1.0)	[1.0, 1.4)	[1.4, 2.4)	[2.4, 5.7]
e-index	[0.0, 3.3]	[0.0, 1.0)	[1.0, 1.4)	[1.4, 1.7)	[1.7, 3.3]
w-index	[0, 8]	[0.0, 1.0)	[1, 2)	[2, 3)	[3, 8]
rational h-index	[0.0, 5.8]	[0.0, 1.0)	[1.0, 1.3)	[1.3, 2.0)	[2.0, 5.8]
impact factor	[0.0, 6.1]	[0.0, 0.4)	[0.4, 0.8)	[0.8, 1.5)	[1.5, 6.1]
immediacy-index	[0.0, 1.2]	[0.0, 0.0]	(0.0, 0.1]	(0.1, 0.2]	(0.2, 1.2]

that the *citations* and the *h-index* are the parent nodes of the *a-index* in the Bayesian network. These nodes are represented in this index's definition. Similarly, the *m-index* definition is also represented in the network, since its parent nodes are the *citations* and the *h-index*. On the other hand, the q^2*-index*, which is dependent on the *h-index* and the *m-index*, has these nodes as parents in the network. The parents of the *r-index* are the *h-index* and the *a-index*. Initially, these indices do not appear in the *r-index* definition, but, after a transformation of the original definition, it is obtained that the *r-index* is also obviously defined as $\sqrt{a \cdot h}$. Finally, the *e-index* is dependent on the *h-index* and the *a-index*. Like the *r-index*, the above indices are not part of the *e-index* definition, but, after few transformations of this definition, the *e-index* can be defined as $\sqrt{a \cdot h - h^2}$.

Besides discovering dependencies between indices, which can be checked against the index definitions, the Bayesian network is also able to discover other kinds of probabilistic dependencies. One example is *h-index* $\rightarrow$ *rational h-index*, not directly derived from, but related to index definitions. In this case, the information about *h-index* influences the probability of the *rational h-index*. Another example (see Figure 7.3) is the probabilistic dependency between *r-index* and *rational h-index*. The objective of the *r-index* is to measure the citation intensity in the *h-core*, whereas the *rational h-index* measures the distance to the next value of the *h-index*. These indices measure different things, but they are probabilistically dependent. More examples of such dependencies are: q^2*-index*$\rightarrow$*w-index*, *r-index*$\rightarrow$*w-index* and q^2*-index*$\rightarrow$*immediacy-index*.

Both the aggregated and the global models have been learned using Elvira software [141]. One of the most useful features of Elvira is the automatic coloring of arcs, which offers qualitative insight about the conditional probability tables attached to each node. This coloring is based on the sign of influence [475] and the magnitude of influence [271].

In order to understand the dependencies between the indices represented in Figure 7.3, some concepts about the influence and color of the arcs are explained. For example, an arc from X to Y is said to have a positive influence if higher values of X lead to higher probabilities of Y taking higher values for any configuration of its other parents. The definition of negative influence and null influence are analogous. When the influence is neither positive

nor negative nor null, then it is said to be undefined. Positive, negative, undefined, and null influence is colored in red, blue, purple, and black, respectively. Taking into account the above concepts, it is observed that *documents* has a positive influence on *citations*. Likewise, *citations* influences the *h-index* and *impact factor* positively. High values of *citations* are associated with high values of *h-index* and *impact factor*. Furthermore, the *h-index* has a positive influence on the *g-index*, which also has a positive influence on the *hg-index*. On the other hand, other arcs in Figure 7.3 represent an undefined influence between the parent and child nodes. Finally, the thickness of the arc is proportional to the magnitude of the influence. Thus, it is found that the relationships that have a grater influence are: *citations*→*h-index*, *h-index*→*g-index* and *g-index*→*hg-index*.

According to the conditional independencies, Table 7.8 lists conditional independencies between indices, derived using the Markov properties in the global Bayesian network. Remember that new conditional independencies can be derived using some conditional independency properties, such as, symmetry, factorization or strong joint. Analyzing Table 7.8, it is observed that some conditional independency relationships are represented in both the aggregated Bayesian network and the global Bayesian network.

It is also observed that other conditional independencies were not shown before because they had a slightly different Bayesian network structure. Note that the *a-index*, the *w-index* and the *rational h-index* are indices whose parents have undergone an important change. Taking the *a-index* independency relationships as an example, it is observed that *citations* and *h-index* are new parents of *a-index* in the global model, and this determines new *a-index* conditional independencies, such as, I(*a-index*, *documents* | *citations*, *h-index*, *g-index*).

The global Bayesian network finds conditional independencies of which some are justified by index definitions. On the other hand, thought, it discovers other conditional independencies that were not derived from such definitions. Looking at the *e-index*, it represents the excess citations received by all papers in the *h-core*. According to this definition, it is reasonable to expect that *e-index* and *citations* would be dependent, but the global model shows that the above indices are independent given *h-index*, *g-index* and *a-index*. Similarly, the relationship between *a-index* and *m-index* is analyzed. The *a-index* is the average number of citations received by the articles included in the *h-core*, whereas the *m-index* is the median number of citations received by papers in the *h-core*. Initially, one might expect there to be a dependency relationship between *a-index* and *m-index*, but the proposed global model suggests that the relationship is of conditional independency, given *citations* and *h-index*. Finally, a *w-index* of at least k means that there are k distinct publications that have at least 1, 2, 3, 4,..., k citations, respectively. According to its definition, the *w-index* should depend on *documents* and *citations*, but the global model shows that they are conditionally independent given q^2*-index*, *r-index* and *e-index*.

7.2.4 Exploiting the global Bayesian network model

It is expected that the best model is the proposed global Bayesian network because its structure reflects more index definitions than the aggregated model and discovers new interesting

Table 7.8: Conditional independencies among indices, derived using Markov properties in the global Bayesian network

Index	is conditionally independent of	given
documents	*h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *w-index*, *rational h-index*, *impact factor*, *immediacy-index*	*citations*
citations	*hg-index*, *e-index*, *w-index*, *rational h-index*, *immediacy-index*	*documents*, *h-index*, *a-index*, *m-index*, *impact factor*
h-index	*immediacy-index*	*citations*, *g-index*, *a-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *rational h-index*
g-index	*w-index*, *rational h-index*	*h-index*, *hg-index*, *a-index*, *e-index*
g-index	*documents*, *citations*, *m-index*, q^2*-index*, *impact factor*, *immediacy-index*	*h-index*
hg-index	*documents*, *citations*, *h-index*, *a-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *w-index*, *rational h-index*, *impact factor*, *immediacy-index*	*g-index*
a-index	*w-index*, *rational h-index*	*citations*, *h-index*, *g-index*, *r-index*, *e-index*
a-index	q^2*-index*, *immediacy-index*	*citations*, *h-index*, *g-index*
m-index	*w-index*, *rational h-index*, *immediacy-index*	*citations*, *h-index*, q^2*-index*
m-index	*hg-index*, *a-index*, *r-index*, *e-index*	*citations*, *h-index*
q^2*-index*	*documents*, *citations*, *hg-index*, *a-index*, *r-index*, *e-index*, *impact factor*	*h-index*, *m-index*
r-index	*documents*, *citations*, *g-index*, *hg-index*, *m-index*, q^2*-index*, *e-index*, *impact factor*, *immediacy-index*	*h-index*, *a-index*
e-index	*documents*, *citations*, *m-index*, q^2*-index*, *rational h-index*, *impact factor*, *immediacy-index*	*h-index*, *g-index*, *a-index*
w-index	*documents*, *citations*, *h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, *rational h-index*, *impact factor*, *immediacy-index*	q^2*-index*, *r-index*, *e-index*
rational h-index	*documents*, *citations*, *g-index*, *hg-index*, *a-index*, *m-index*, *e-index*, *w-index*, *impact factor*, *immediacy-index*	*h-index*, q^2*-index*, *r-index*
impact factor	*h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *w-index*, *rational h-index*	*citations*
immediacy-index	*documents*, *citations*, *h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, q^2*-index*, *r-index*, *e-index*, *w-index*, *rational h-index*	*impact factor*

conditional independencies between indices. For this reason, evidence propagation and abduction are applied to the proposed global model.

So far, the graphical component of global Bayesian network has been used to discover conditional (in)dependencies. In this section, the probabilistic component of the global Bayesian network is also used to precisely quantify, the effect of knowing some fixed variables on the

occurrence of other variables. In this context, evidence propagation usually refers to computing the posterior probability of each single variable given the available evidence (i.e., some fixed variables), while abduction consists of finding the most probable configuration of a set of variables of interest given the evidence.

As regards evidence propagation, it would like to know the effect on index probabilities of introducing some specific values for other indices as evidence. The first inference is to fix the *citations* value to *medium-low*. After setting this evidence level, the posterior probabilities of each index are calculated in Figure 7.4, top. Note that the mode (green bars) of all indices is *low* or *medium-low*. Taking the *impact factor* as an example, the following probabilities are observed:

- P(*impact factor=low* | *citations=medium-low*)=0.37
- P(*impact factor=medium-low* | *citations=medium-low*)=0.30
- P(*impact factor=medium-high* | *citations=medium-low*)=0.22
- P(*impact factor=high* | *citations=medium-low*)=0.11

These conditional probabilities are reasonable since fixing *citations=medium-low* as evidence, *impact factor*, which depends on *citations* (positive influence), the value of the mode should be *low* or *medium-low*. Analyzing the above conditional probabilities, it is found that *low* and *medium-low* are the most probable values, at 0.37 and 0.30, respectively.

On the other hand, the second inference is to assign a *high* value to *citations*, see Figure 7.4, bottom. In this case, the value of the mode of most of the indices is *high*. Now, these conditional probabilities are also reasonable and the probability values of the *impact factor* are:

- P(*impact factor=low* | *citations=high*)=0.02
- P(*impact factor=medium-low* | *citations=high*)=0.10
- P(*impact factor=medium-high* | *citations=high*)=0.31
- P(*impact factor=high* | *citations=high*)=0.56

The probabilities of the above inferences answer a question raised in the introduction, namely, what would happen to a specific journal impact factor if the papers, that it published received more citations?. The answer lies in the total distribution of the different impact factor values. Similarly, setting *citations=low*, it is answered the question, what would happen to a specific journal impact factor if the papers that it published received fewer citations?.

To get the most likely plausible explanation *P(configuration | evidence)*, it should be searched the configuration of values of the non-observed indices (called explanation set) that maximizes the above probability. This is possible using abductive inference [357].

Table 7.9 shows three examples of abductive inference. Three different evidence levels are set at *h-index=medium-low* (like, e.g., International Journal of Pattern Recognition and Artificial Intelligence), *impact factor=medium-high* (like, e.g., International Journal of Intelligent

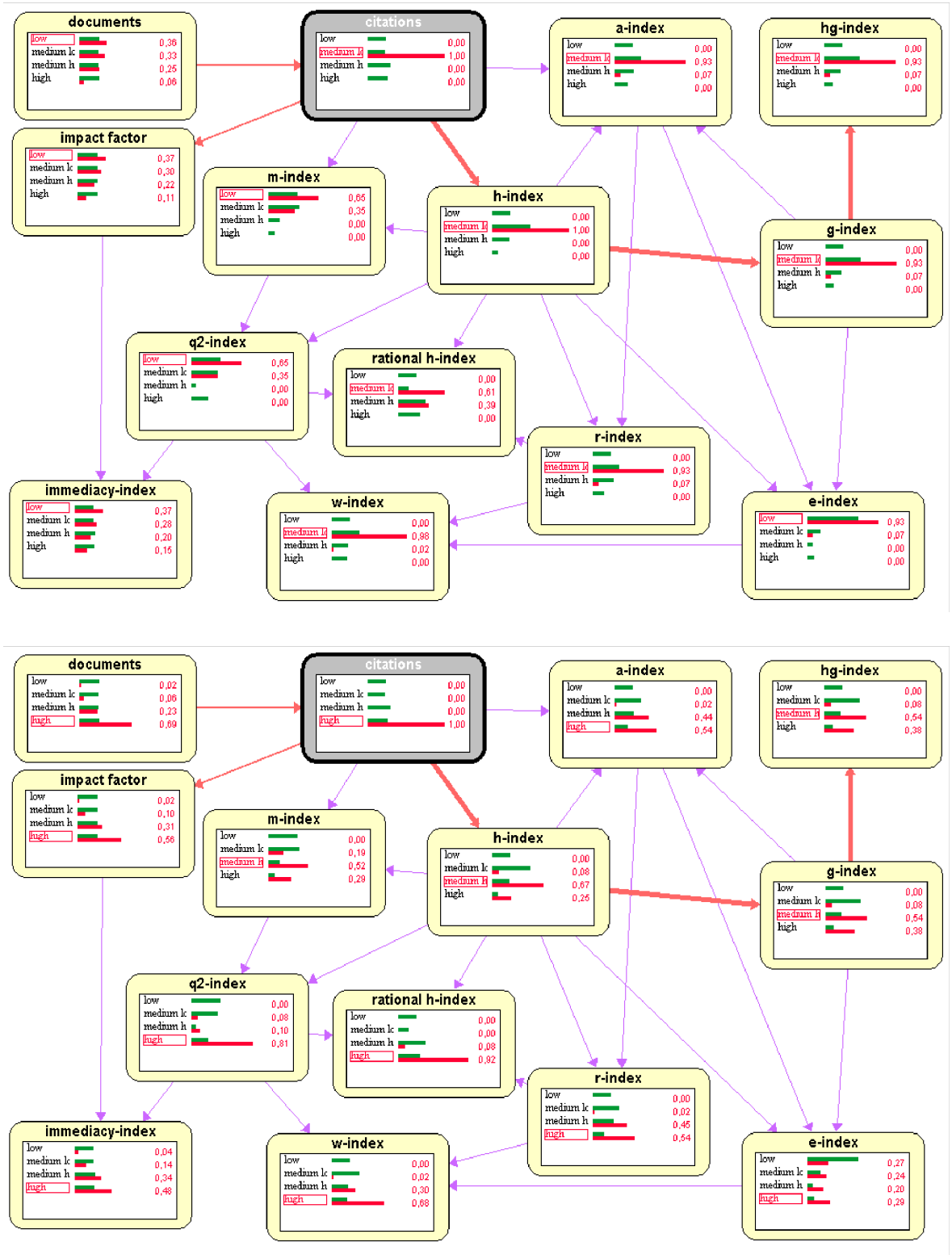


Figure 7.4: Index probabilities after setting the *citations* value at *medium-low* (top) and *high* (bottom)

Table 7.9: Most likely configurations of indices for a given evidence level

Explanation set	Evidence 1	Evidence 2	Evidence 3
	h-index=	*impact factor=*	*immediacy-index=*
Index (X_i)	*medium-low*	*medium-high*	*high*
documents	*low*	*medium-high*	*high*
citations	*medium-low*	*medium-high*	*high*
h-index	-	*medium-low*	*medium-high*
g-index	*medium-low*	*medium-low*	*medium-high*
hg-index	*medium-low*	*medium-low*	*medium-high*
a-index	*medium-low*	*medium-low*	*medium-high*
m-index	*low*	*medium-low*	*medium-high*
q^2*-index*	*low*	*medium-low*	*high*
r-index	*medium-low*	*medium-low*	*medium-high*
e-index	*low*	*low*	*low*
w-index	*medium-low*	*medium-low*	*medium-high*
rational h-index	*medium-low*	*medium-high*	*high*
impact factor	*low*	-	*high*
immediacy-index	*low*	*medium-high*	-
P(*Explanation set* | *evidence*)	0.000245	0.000147	0.000839

Systems) and *immediacy-index=high* (like, e.g., Machine Learning) in Table 7.9, columns 2, 3, and 4, respectively. Taking the third inference as an example, it is showed that the most probable configuration of index values when *immediacy-index=high* is *documents=high*, *citations=high*, *h-index=medium-high*, *g-index=medium-high*, *hg-index=medium-high*, *a-index =medium-high*, *m-index=medium-high*, q^2*-index=high*, *r-index=medium-high*, *e-index=low*, *w-index=medium-high*, *rational h-index=high* and *impact factor=high*. The above configuration of index values answers the question, what kind of actions should journal editorial boards take to get a higher *immediacy-index* value in a specific year?. The answer is publish a lot of *documents* (≥ 79), receive a lot of *citations* (≥ 12), get high q^2*-index* (≥ 2.0), *rational h-index* (≥ 2.0) and *impact factor* (≥ 1.5) index values. Moreover, journal editorial boards should aspire to the following index values: *h-index*=[2,3), *g-index*=[2,3), *hg-index*=[1.4,2.4), *a-index*=[2.0,3.0), *m-index*=[2.0,3.0), *r-index*=[1.4,2.4), *e-index*=[0.0,3.3) and *w-index*=[2,3). Finally, although the joint probability (P=0.000839) of these index values seems very low, the number of different configurations of index values is 4^{13}. Therefore, the joint probability obtained via abduction is considerably greater than would be expected purely by chance ($\frac{1}{4^{13}} = 1.5 \cdot 10^{-8}$).

7.3 Discussion and conclusions

Bibliometric indices have received a lot of attention from the scientific community over the last few years since they are used to evaluate the importance of research at different levels by funding agencies and promotion committees. In view of the vast number of bibliometric indices, it is necessary to analyze how they relate to each other (irrelevant, dependent and so on).

A case study of 14 well-known bibliometric indices on computer science and artificial intelligence journals was performed. Several Bayesian network models were learned from data

to analyze the relationships among bibliometric indices. The induced Bayesian networks are then used to discover probabilistic conditional (in)dependencies among the indices and, also for probabilistic reasoning. The aim of these models is to represent relationships between index values using a within one-year publication and citation window.

Analyzing the best proposed Bayesian network, it is observed that its structure matches many index definitions. In addition, this model learns new knowledge derived from index definitions and discovers new interesting conditional (in)dependencies between analyzed indices. These conditional (in)dependency relationships have been analyzed using Markov properties. Using the proposed models, editorial boards of journals could find the answer to questions related to their journal citation indices. Evidence propagation and abduction inference in Bayesian networks are very useful for answering bibliometric questions.

In the future, the target will be to build new models that incorporate other journal citation indices like *eigenfactor*, *article influence* and *Scimago's journal rank index*, among others. These models could also be induced using different Bayesian network learning algorithms. The way index values are handled influences the results. They could be modeled as continuous variables instead of discretizing the values.

Part IV

Exploring Spanish Computer Science Research

Dataset compilation

The purpose of Part IV is to analyze the computer science research in Spain. To do this, a bibliometric dataset is built containing the research activity of Spanish academic staff in the computer science area. The first phase of the dataset compilation was to apply to the Spanish Ministry of Education for a list of academics associated with the computer science area who were active as of January 1, 2010. This list includes the full name of 2004 academics, and their associated university, position and research area. All the academic staff are associated with one of the following three specific areas: Computer architecture and technology (CAT), computer science and artificial intelligence (CSAI), and computer languages and systems (CLS). Members of the academic staff specialize in one of the above specific area, in which they lecture, regularly publish and are assessed by national organizations. There are four types of permanent (civil servant) positions in the Spanish higher education system. All four positions are associated with tenure obligations. These academic staff work full-time and engage in teaching, mentoring and research at the university. These four academic positions are translated (from the highest to the lowest level) as: full professor (FP), associate professor-type1 (AP1), associate professor-type2 (AP2) and associate professor-type3 (AP3).

The next step was to retrieve a list of publications and citation data from the date of the first publication by an academic staff (January 1, 1973) to January 1, 2010. This information was carefully downloaded from the *Web of Science* research platform bearing in mind Spanish personal name variations in international databases.

Web of Science is selected as bibliographic database because it is the most comprehensive and versatile research platform available and has an important reputation as the oldest citation resources, containing the most prestigious academic journals. It also records a very large part of scientific literature and what really matters. Moreover, it is one of the most important tools used by CNEAI and ANECA in order to assess Spanish scientific activity. According to the computer science area, *Web of Science* contains databases specialized in journals and conferences, indexing more than 470 computer science journals and more than 15,000 of the major computer science conferences.

Regarding data extraction, only documents considered as journal articles and conference papers were taken into consideration. Also, the publication subject classification was used as a filter. In this way, only documents published in journals and conferences belonging to the seven major fields of computer science were taken into account. According to the *Journal Citation Reports* these major fields are: artificial intelligence, cybernetics, hardware and architecture, information systems, interdisciplinary applications, software engineering and theory and methods. Finally, in order to ensure the reliability of results, the final list of publications were checked against other databases like DBLP Computer Science Bibliography, personal webpages and institutional websites, among others. The last phase was to develop a software which used all this information in order to calculate bibliometric indices. Different measures are used according to the specific analysis carried out in the following chapters.

Chapter 8

Overview of Spanish Computer Science Research

8.1 Introduction

Scientific production is a crucial element to evaluate the research activity of a country. The national production is usually published by higher education institutions which play an important role in national research. Spanish higher education has expanded remarkably over the last half century. There are now 50 public universities and 28 private universities, compared with only 15 in 1968. Most of these institutions sprang up between the 1970s and the 1990s. Because of this rapid spread, research has grown exponentially in Spain over last years. It now accounts for 3.3% of global output stored in the *Web of Science* research platform, compared with a share of only 0.2% in 1963. But, as in other areas, quality is more important than quantity in science. And this is where Spain falls down. According to *Essential Science Indicators*, Spain ranks 9th among the top-performing countries for papers, 11th for citations, and 39th for citations per paper in all fields.

Spain fares no better in the field of computer science. On the one hand, Spain now ranks 9th for papers (14,904 publications). The top five countries in this ranking are USA (81,475 publications), China (43,899 publications), Germany (20,094 publications), England (19,330 publications) and South Korea (18,757 publications). On the other hand, Spain ranks 7th for citations (63,293 citations) behind countries such as USA (596,994 citations), China (150,885 citations), England (122,176 citations), Germany (120,066 citations), and France (89,037 citations). These rankings show that countries that rank the top positions for papers, usually rank the top positions for citations. Finally, Spain ranks 30th for citations per paper (4.25). In this case, Switzerland (9.76) Ireland (7.94) and USA (7.33) are the top-performing countries.

The Spanish scientific production has been analyzed in many areas such as medicine [183], communications [298], information technology [381], philology [443], mathematics [444], etc. Other studies [49, 185, 335] have analyzed the number of documents, citations and patents

published by Spanish universities. Finally, several university research profiles [76, 336, 437] and regions research profiles [348, 377] have been also analyzed in Spain.

The computer science area has become as one of the main drivers of economic growth in recent years, achieving great progress and bridging the gap between the university and the business world. Some works [187, 189] have analyzed this area worldwide, whereas others works [202, 466] have focused on specific countries. In contrast, an analysis of the computer science in Spain has never carried out. An overview of the Spanish computer science research allows to assess the performance of the scientific activity and its impact on the society. In this context, a bibliometric analysis could provide information about the Spanish trends in the sector and the research profiles of Spanish universities and departments.

The aim of this chapter is to illustrate the productivity and visibility of Spanish computer science research. To do this, research produced in higher education institutions and their academic staff are taken into account. This analysis is confined to public universities, and also circumscribed to 48 out of the 50 public universities, because two of them (Universidad Internacional de Andalucía and Universidad Internacional Menéndez Pelayo) have no academic staff specialized in the computer science field. Regarding the academic staff, the 2004 tenured academics are analyzed according to their area (computer architecture and technology (CAT), computer science and artificial intelligence (CSAI), and computer languages and systems (CLS)) and position (full professor (FP), associate professor-type1 (AP1), associate professor-type2 (AP2) and associate professor-type3 (AP3)).

Finally, the Spanish computer science research is analyzed by different parameters, such as, number of documents, number of citations, number of citations per document, number of authors per document, number of institutions per document, document types, types of collaboration, computer science disciplines, impact factor and some bibliometric indices. The results are presented at different levels of detail (nationwide, autonomous regions, public universities, subject areas and professional standing). Thanks to this chapter a comprehensive overview of the current situation in the area of computer sciences is achieved. This chapter is based on the publications [219, 222].

Chapter outline

Section 8.2 presents an overview of the Spanish computer science research, including different analysis at the macro level (Section 8.2.1), meso level (Sections 8.2.2 and 8.2.3) and micro level (Section 8.2.4). Finally, conclusions are discussed in Section 8.3.

8.2 Analyzing Spanish computer science

Different parameters are computed to get an overview of the Spanish computer science research. The number of documents and the number of citations are the basic measures for representing the productivity and visibility, respectively, whereas the number of citations per document represents the quality of publications. Regarding collaboration, the number of au-

thors per document and the number of institutions per documents could be used as measures of collaboration's intensity. The trend of number of documents, citations and citations per document are also analyzed according to the document type, that is, documents published as journal articles and proceeding papers. This analysis also studies productivity and visibility according to different computer science disciplines. Furthermore, this analysis explores how the number of documents and citations vary across national and international collaboration. Focusing on the impact factor, the number of documents published in different journal impact factor quartiles is also analyzed. Finally, some well-know bibliometric indices are shown.

In favor of the repeatability principle, some general notes of the data computation process are shown: a) All publications are stored in several database tables with the intention of exploiting the information at different levels (nationwide, autonomous regions, public universities, subject areas and professional standing). Taking a document published by researchers affiliated to universities A and B (both belong to the same autonomous region) as an example, two records of the above document are stored in the public universities table, whereas only one record is stored in the nationwide table and autonomous regions table. In this way, the overlap produced by the collaboration among researchers is removed; b) Regarding the impact factor associated with each publication, only values from the corresponding *Journal Citation Reports* edition during the period (2000-2009) were extracted. Also, a journal could belong to different quartiles (Q1, Q2, Q3, Q4) according to the selected discipline (there is some overlap among disciplines, that is, some journals could belong to different disciplines). In this way, the best quartile value for each journal is always selected.

8.2.1 Nationwide results

Table 8.1 shows an overview of the Spanish computer science research. Results are shown using different parameters such as productivity, visibility, authorship, document types, types of collaboration, computer science disciplines, impact factor and some bibliometric indices.

The research productivity of the Spanish academics represents 11,510 publications. These publications can be divided into 4,233 journal articles (36.8%) and 7,277 proceedings papers (63.2%). Most of documents are usually published by Theory and Methods and Artificial Intelligence disciplines. Regarding the types of collaboration, only 1,601 publications (13.9%) include an international institution as collaborator. Focusing on the journal impact factor, most of journal articles are published in first-quartile of journal impact factor during the 2000-2009 period, achieving percentages of Q1 (30.5%), Q2 (27.5%), Q3 (27.3%), and Q4 (14.7%), respectively.

The Spanish research visibility represents 37,333 citations received by all published documents, achieving 3.24 citations per document. The computer science discipline which receives more citations is Artificial Intelligence (15,907 citations), whereas Cybernetics is the discipline which receives more citations per document (5.80). Regarding the type of publication, journal articles receives 7.18 citations per document, whereas proceedings papers only receive 0.96 citations per document. Finally, documents published with international collaborators receive more citations (5.54 citations per document) than documents published with national

Table 8.1: Scientific production of Spanish academics belonging to the computer science area.

General data

Number of researchers: 2,004	3.24 citations per publication
Number of publications: 11,510	3.54 authors per publication
Number of citations: 37,333	2.39 institutions per publication

Document type

	Publications	Citations	Rate
Journal article	4,233	30,378	7.18
Proceedings paper	7,277	6,955	0.96

Types of collaboration

	Publications	Citations	Rate
National collaboration	9,909	28,471	2.87
International collaboration	1,601	8,862	5.54

JCR categories

	Publications	Citations	Rate
C.S. Artificial Intelligence	3,667	15,907	4.34
C.S. Cybernetics	238	1,381	5.80
C.S. Hardware and Architecture	1,075	3,112	2.89
C.S. Information Systems	1,406	3,843	2.73
C.S. Interdisciplinary Applications	818	2,013	2.46
C.S. Software Engineering	1,632	3,943	2.42
C.S. Theory and Methods	5,921	15,481	2.61

Bibliometric indices

h-index: 63	*g-index*: 92	*hg-index*: 76.1
h_{pub}*-index*: 45	h_{cit}*-index*: 119	h_h*-index*: 13

Impact factor

976 articles published by Q1 journals	879 articles published by Q2 journals
875 articles published by Q3 journals	471 articles published by Q4 journals

collaborators (2.87 citations per document).

Table 8.1 also shows information about bibliometric indices. The *h-index*, *g-index*, and *hg-index* values are 63, 92 and 76.13, respectively. Other bibliometric indices values are: h_{pub}-*index*=46, h_{cit}-*index*=122 and h_h-*index*=13. These values mean that 46 Spanish academics have published at least 46 publications, 122 academics have received at least 122 citations, and 13 academics have at least *h-index*=13.

Table 8.2 represents the evolution of the above parameters during the period 2000-2009. Spanish research productivity and visibility have increased their values in the last years, achieving an increment of 347% and 1,053%, respectively. Regarding collaboration intensity, the average number of authors per publication was 3.12 in 2000 and 3.72 in 2009, whereas the average number of institutions per document have fluctuated during the analyzed period.

Regarding the type of publications, Table 8.3 lists the name of journals and conferences that have published the most of the documents belonging to the Spanish academic staff. By journals, *Fuzzy Sets and Systems* published the highest number of Spanish publication (126 documents) during 2000-2009. It published 15 documents in 2009, whereas *Expert Systems*

Table 8.2: Evolution of the scientific production and visibility of Spanish academics belonging to the computer science area.

General data	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Total publications	424	480	589	996	1,001	1,100	1,224	1,300	1,337	1,471
Total citations	662	997	1,251	1,876	2,389	2,990	3,595	4,418	6,136	6,971
Average authors	3.12	3.20	3.37	3.53	3.58	3.62	3.64	3.69	3.74	3.72
Average institutions	2.35	2.45	2.52	2.57	2.60	2.53	2.55	2.30	2.30	2.37
Document type										
Journal articles	174	215	226	265	276	286	335	419	456	549
Proceeding papers	250	265	363	731	725	814	889	881	881	922
Cits in journal articles	614	902	1,119	1,581	1,928	2,403	2,809	3,423	4,888	5,527
Cits in proceeding papers	48	95	132	295	461	587	786	995	1,248	1,444
Types of collaboration										
National publications	377	423	506	827	856	896	1,011	1,174	1,201	1,328
International publications	47	57	83	169	145	204	213	126	136	143
Cits in national public.	529	777	974	1,432	1,821	2,276	2,748	3,341	4,659	5,306
Cits in international public.	133	220	277	444	586	714	847	1,077	1,477	1,665
JCR categories										
Artificial Intelligence	187	189	232	299	297	309	384	393	424	415
Cybernetics	4	10	15	12	19	10	40	25	17	29
Hardware and Architecture	37	58	71	58	77	73	105	100	148	111
Information Systems	43	81	79	92	101	105	122	168	184	211
Interdisciplinary Applications	15	27	24	44	54	68	53	117	104	148
Software Engineering	48	85	80	105	132	103	130	219	238	223
Theory and Methods	162	190	249	580	563	675	663	699	654	799
Bibliometric indices										
h-index	20	22	24	28	33	37	43	49	55	60
g-index	27	30	34	38	43	49	57	65	77	86
hg-index	23.2	25.7	28.6	32.6	37.7	42.6	49.5	56.4	65.1	71.8
Impact factor										
Publications in Q1 journals	34	57	51	58	76	82	108	128	130	252
Publications in Q2 journals	30	33	36	63	84	81	108	134	131	179
Publications in Q3 journals	73	88	66	109	88	80	81	105	110	75
Publications in Q4 journals	37	37	73	35	28	43	38	52	85	43

with Applications published 35 documents in 2009. Regarding conferences, the *International Conference on Artificial Neural Networks*, although it is a biannual conference, published the highest number of proceedings papers (280 documents).

Table 8.2 also illustrates that the number of proceedings papers is higher than the number of journal articles in each analyzed year. Despite this, the number of citations received by journal articles is much higher than the number of citations received by proceedings papers. Taking the 2009 year as an example, journal articles received 5,527 citations whereas proceedings papers received 1,444 citations. Figure 8.1 shows the evolution of the average number of citations received by journal articles and proceeding papers. Taking the 2000 year

Table 8.3: Number of documents published by journals and conferences that have the top number of Spanish publications during the period 2000-2009.

Journals	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Fuzzy Sets and Systems	14	15	7	14	16	11	5	16	13	15
Int. Journal of Intelligent Systems	10	5	6	21	3	12	7	8	9	9
Expert Systems with Applications	5	2	4	4	7	6	3	5	14	35
Pattern Recognition Letters	3	7	4	10	2	5	8	9	18	7
Pattern Recognition	6	3	9	10	6	5	13	4	8	7
Conferences	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Artificial Neural Networks	0	4	0	72	0	56	0	43	0	105
Cross-Language Evaluation Forum	0	0	5	21	0	15	25	23	22	18
Natural and Artificial Computation	0	0	0	0	0	31	0	32	0	34
Computational Science	0	0	11	11	0	11	23	4	15	11
Computer Aided Systems Theory	0	0	0	0	0	32	0	34	0	17

as an example, journal articles published in 2000 had an average value of 15.24 citations per document in 2010, whereas proceedings papers published in 2000 had 1.61 citations per document in 2010.

Regarding the type of collaboration, the number of national collaboration has increased each analyzed year. In contrast, the number of international collaboration has increased during the first years, and then, has decreased the last three years. By *JCR* categories, Artificial Intelligence and Theory and Methods are the disciplines with highest publication in each year. Table 8.2 also shows the evolution of the bibliometric indices values. These values have increased during the analyzed period. Taking the *h-index* as an example, its value was 20 in 2000, whereas it was 60 in 2009. Since these bibliometric indices values never decreased, it is important to analyze the increase of these values year-by-year. In this context, the increment achieved during the first years was lower than four units, whereas the increment achieved during the last years was higher than four units.

Table 8.2 also presents the evolution of documents published in each impact factor quartile. The number of documents published in Q1 and Q2 journals have increased during the analyzed period. Note that 34 and 30 documents were published in 2000 by Q1 and Q2 journals, whereas 252 and 179 documents were published in 2009 by Q1 and Q2 journals. Figure 8.2 shows the percentage of documents published by each impact factor quartile. Taking Q1 and Q4 quartiles as examples, the year with the highest percentage of Q1 documents was 2009, whereas the year with the highest percentage of Q4 documents was 2002. In this context, the percentages of Q1 and Q2 documents have increased during the analyzed period, whereas the percentages of Q3 and Q4 documents have decreased. Figure 8.3 shows the number of times that documents are published by journals with a specific impact factor during 2000-2009. The distribution of the impact factor values is: minimum (0.000), percentile 25 (0.470), percentile 50 (0.799), percentile 75 (1.282), and maximum (7.400). Furthermore, the most frequent impact factor is 2.596.

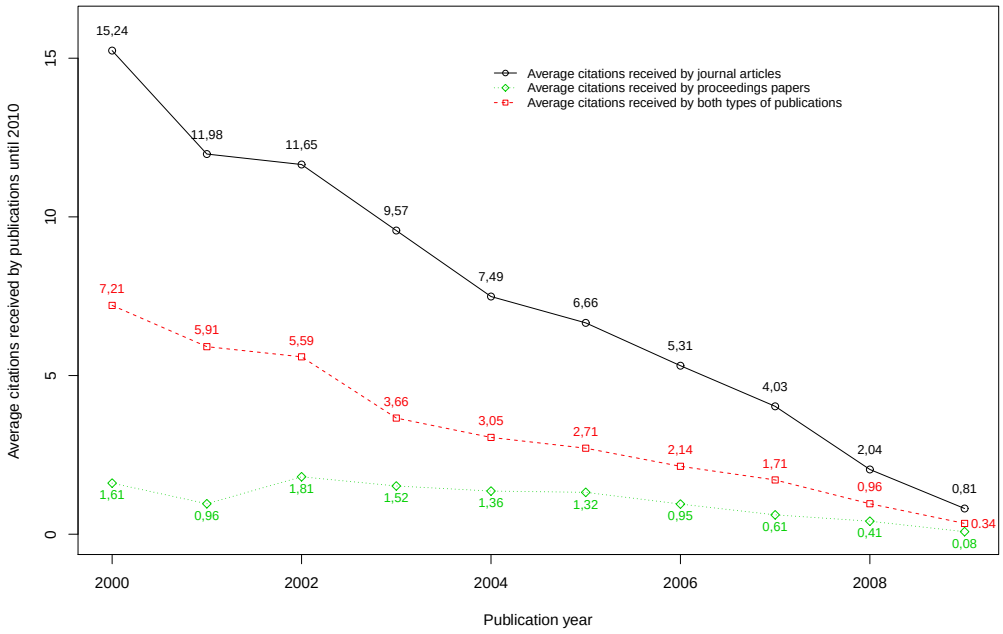


Figure 8.1: Evolution of the average number of citations received by scientific publications (journal articles and proceeding papers).

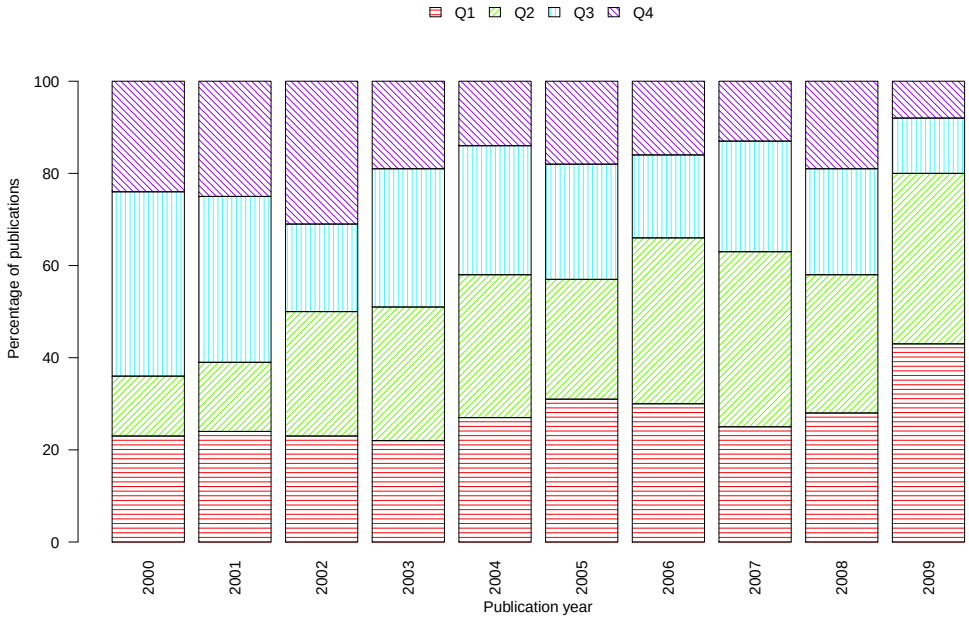


Figure 8.2: Evolution of the percentage of documents published by each impact factor quartile.

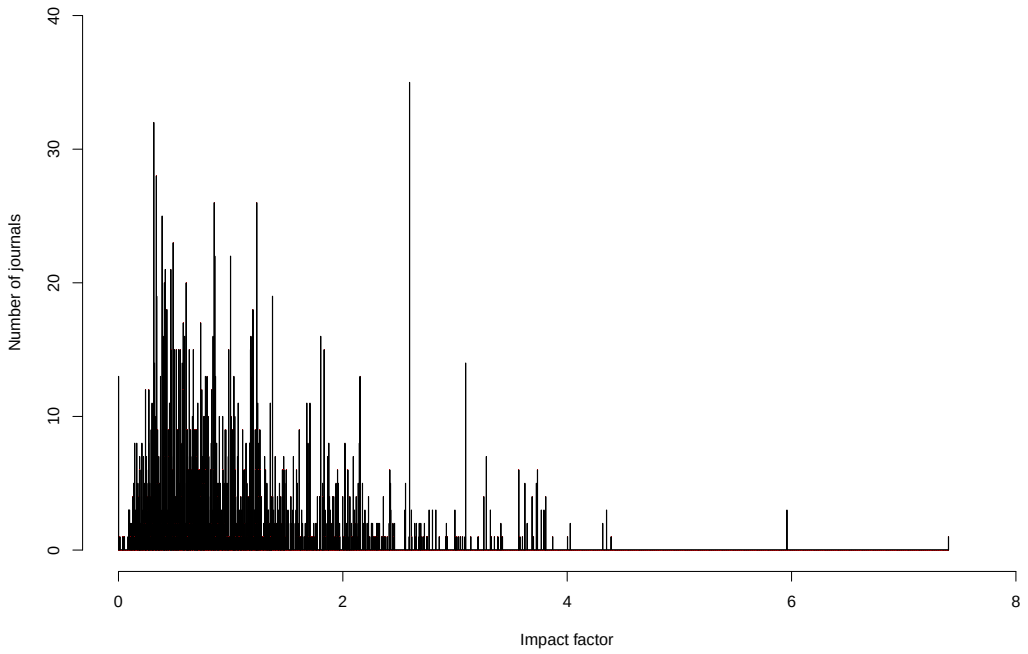


Figure 8.3: Number of times that documents are published by journals with a specific impact factor.

Finally, some relationships among several parameters are also analyzed. On the one hand, the relationships among publications, citations, *JCR* disciplines and document types are explored. Table 8.4 (top) shows that Artificial Intelligence is the discipline which publishes the highest number of journal articles, whereas Theory and Methods is the discipline which publishes the highest number of proceedings papers. Also, the 66.4% of documents published in Cybernetics are journal articles, whereas the 79.9% of documents published in Theory and Methods are proceedings papers. Regarding citation values, Artificial Intelligence and Theory and Methods categories receive the highest number of citations. Focusing on the percentage of citations received by journal articles and proceedings papers, the 1.0% of citations received by Cybernetics category correspond to proceedings papers. In contrast, the 35.1% of citations received by Theory and Methods correspond to proceedings papers. On the other hand, it is also explored how publications, citations, *JCR* disciplines and collaboration types are related. Table 8.4 (down) shows that the percentages of documents published in national collaboration range between 83.3% and 90.8%. The *JCR* categories which have the highest percentages of international collaborations are Software Engineering (16.7%), Hardware and Architecture (14.7%) and Information Systems (14.4%). By visibility, the 12.8% of citations received by Cybernetics documents correspond to international collaborations, whereas the 35.3% of citations received by Software Engineering documents correspond to international collaborations.

Table 8.4: Relationships among publications, citations, document types (journal articles and proceeding papers) and collaboration types (national and international) by *JCR* categories.

JCR categories	Publications		Citations	
	Articles	Proceedings	Articles	Proceedings
Artificial Intelligence	1,760 (48.0%)	1,909 (52.0%)	14,764 (92.8%)	1,143 (7.2%)
Cybernetics	158 (66.4%)	80 (33.6%)	1,367 (99.0%)	14 (1.0%)
Hardware and Architecture	471 (43.8%)	604 (56.2%)	2,765 (88.8%)	347 (11.2%)
Information Systems	662 (47.1%)	774 (52.9%)	3,571 (92.9%)	272 (7.1%)
Interdisciplinary Applications	433 (52.9%)	385 (47.1%)	1,919 (95.3%)	94 (4.7%)
Software Engineering	829 (50.8%)	803 (49.2%)	3,705 (94.0%)	238 (6.0%)
Theory and Methods	1,189 (20.1%)	4,732 (79.9%)	10,047 (64.9%)	5,434 (35.1%)
	National	International	National	International
Artificial Intelligence	3,274 (89.3%)	393 (10.7%)	12,948 (81.4%)	2,959 (18.6%)
Cybernetics	216 (90.8%)	22 (9.2%)	1,204 (87.2%)	177 (12.8%)
Hardware and Architecture	917 (85.3%)	158 (14.7%)	2,024 (65.0%)	1,088 (35.0%)
Information Systems	1,203 (85.6%)	203 (14.4%)	2,823 (73.5%)	1,020 (26.5%)
Interdisciplinary Applications	727 (88.9%)	91 (11.1%)	1,494 (74.2%)	519 (25.8%)
Software Engineering	1,359 (83.3%)	273 (16.7%)	2,552 (64.7%)	1,391 (35.3%)
Theory and Methods	5,116 (86.4%)	805 (13.6%)	11,697 (75.6%)	3,794 (24.4%)

8.2.2 Autonomous region results

Table 8.5 shows the scientific production by autonomous regions. The number of academic staff (A), the number of publications (P), the number of citations (C), the number of citations per publication (C/P), the number of publications per academic (P/A), the number of citations per academic (C/A), the percentage of documents published by journals (JRN), the percentage of documents published by Q1 journals (Q1), the percentage of documents with international collaboration (COL), and the h-index value (H) are calculated for each autonomous region. Analyzing the above parameter values, some conclusions are presented:

- The autonomous regions with the highest number of academic staff are: Madrid (383), Andalucía (353), Valencia (352), Cataluña (261) y Galicia (94). These values have an important influence in the fact that most of these autonomous regions have the highest number of publications: Andalucía (2,410), Madrid (2,258), Cataluña (2,009), Valencia (1,929), and C. Mancha (594).

- A higher number of academics also affects the number of citations. Thus, the autonomous regions with the highest number of citations are: Andalucía (13,421), Cataluña (6,592), Madrid (5,747), Valencia (5,098) and País Vasco (1,311). This ranking changes when the ratio between citations and publications is calculated: Navarra (6.1), Andalucía (5.6), País Vasco (4.0), Islas Baleares (3.8) and Cataluña (3.3).

- The autonomous regions with the highest number of publications per academic are: Cantabria (9.0), C.Mancha (8.9), Cataluña (7.7), Navarra (7.6) and Andalucía (6.8).

Table 8.5: Scientific production of Spanish autonomous regions. Numbers in bold represent the top five positions in each parameter.

Aut.Region	A	P	C	P/C	P/A	C/A	JRN%	Q1%	COL%	H
Andalucía	**353**	**2,410**	**13,421**	**5.6**	**6.8**	**38.0**	**47.2**	28.3	**53.6**	**52**
Aragón	38	165	508	3.1	4.3	13.4	36.4	**37.0**	42.4	11
Asturias	51	181	527	2.9	3.5	10.3	**43.1**	33.3	39.0	11
Canarias	75	213	229	1.1	2.8	3.1	22.1	21.6	28.3	8
Cantabria	13	117	350	3.0	**9.0**	**26.9**	**43.6**	24.2	**64.3**	11
C. León	73	274	554	2.0	3.8	7.6	31.4	29.6	42.3	12
C. Mancha	67	594	1,085	1.8	**8.9**	16.2	35.0	23.4	42.9	16
Cataluña	**261**	**2,009**	**6,592**	**3.3**	**7.7**	**25.3**	38.4	**36.6**	**64.8**	**33**
Extremadura	46	112	194	1.7	2.4	4.2	33.0	18.2	36.1	8
Galicia	**94**	539	1,275	2.4	5.7	13.6	37.7	**36.5**	32.9	15
Islas Baleares	47	204	766	**3.8**	4.3	16.3	**50.5**	29.9	40.0	14
La Rioja	6	21	56	2.7	3.5	9.3	38.1	0.0	**54.5**	5
Madrid	**383**	**2,258**	**5,747**	2.5	5.9	15.0	34.7	29.9	49.7	**27**
Murcia	61	303	706	2.3	5.0	11.6	39.6	**36.2**	28.1	12
Navarra	15	114	700	**6.1**	**7.6**	**46.7**	**64.0**	26.7	23.4	15
País Vasco	69	324	**1,311**	**4.0**	4.7	**19.0**	39.2	**34.9**	33.8	17
Valencia	**352**	**1,929**	**5,098**	2.6	5.5	14.5	29.5	31.8	**70.9**	**28**

In the same context, the ranking of autonomous regions with the highest number of citations per academic are: Navarra (46.7), Andalucía (38.0), Cantabria (26.9), Cataluña (25.3) and País Vasco (19.0).

- Regarding the type of publication, the autonomous regions have different behaviors. Canarias and Valencia usually publish in conferences. The 70% of their publications are proceedings papers. In contrast, autonomous regions like Navarra and Islas Baleares are the ones that publish more documents in journals than in conferences. The 64.0% and 50.5% of their publications are journal articles, respectively.

- The autonomous regions with the highest percentage of documents published by Q1 journals are: Aragón (37.0%), Cataluña (36.6%), Galicia (36.5%), Murcia (36.2%) and País Vasco (34.9%).

- Regarding international collaboration, the top positions in this ranking are: Valencia (70.9%), Cataluña (64.8%), Cantabria (64.3%), La Rioja (54.5%) and Andalucía (53.6%).

- Analyzing the h-index value, the autonomous regions with the highest values are Andalucía (52), Cataluña (33), Valencia (28), Madrid (27) and País Vasco (17).

8.2.3 University results

Table 8.6 shows the scientific production by Spanish public universities. The number of academic staff (A), the number of publications (P), the number of citations (C), the number of citations per publication (C/P), the number of publications per academic (P/A), the number

of citations per academic (C/A), the percentage of documents published by journals (JRN), the percentage of documents published by Q1 journals (Q1), the percentage of documents with international collaboration (COL), and the h-index value (H) are calculated for each university. Analyzing the above parameter values, some conclusions are presented:

Table 8.6: Scientific production of Spanish public universities. Numbers in bold represent the top five positions in each parameter.

University	A	P	C	P/C	P/A	C/A	JRN%	Q1%	COL%	H
A Coruña	45	325	484	1.5	7.2	10.8	29.8	28.2	25.2	11
Alcalá	41	112	193	1.7	2.7	4.7	53.6	26.5	57.1	7
Alicante	77	392	841	2.1	5.1	10.9	25.8	38.6	60.0	13
Almería	37	119	186	1.6	3.2	5.0	31.9	20.0	25.0	7
Aut. Barcelona	46	368	716	1.9	8.0	15.6	25.8	36.9	51.6	10
Aut. Madrid	27	253	773	3.1	**9.4**	28.6	37.9	44.0	58.8	12
Barcelona	7	52	74	1.4	7.4	10.6	48.1	**52.6**	19.3	4
Burgos	9	33	125	3.8	3.7	13.9	30.3	**57.1**	54.6	6
Cádiz	32	49	62	1.3	1.5	1.9	30.6	7.1	14.3	4
Cantabria	13	117	350	3.0	9.0	26.9	43.6	24.2	**64.3**	11
Carlos III	37	395	494	1.3	**10.7**	13.4	30.9	23.8	29.6	10
C. Mancha	67	594	1,085	1.8	8.9	16.2	35.0	23.4	42.9	16
Complutense	65	612	1,607	2.6	**9.4**	24.7	30.9	22.6	51.0	15
Córdoba	20	92	436	4.7	4.6	21.8	**66.3**	**67.3**	31.0	12
Extremadura	46	112	194	1.7	2.4	4.2	33.0	18.2	36.1	8
Girona	27	170	863	**5.1**	6.3	32.0	44.7	32.8	**79.3**	14
Granada	**93**	**1,107**	**9,882**	**8.9**	**11.9**	**106.3**	**59.5**	28.8	31.5	**50**
Huelva	8	17	38	2.2	2.1	4.8	35.3	20.0	0.0	4
Illes Balears	47	204	766	3.8	4.3	16.3	50.5	29.9	40.0	14
Jaén	27	156	1,483	**9.5**	5.8	**54.9**	50.6	24.6	21.8	17
Jaume I	58	375	971	2.6	6.5	16.7	32.5	26.4	**61.1**	15
La Laguna	17	116	156	1.3	6.8	9.2	24.1	4.0	32.1	6
La Rioja	6	21	56	2.7	3.5	9.3	38.1	0.0	54.6	5
Las Palmas GC	58	97	73	0.8	1.7	1.3	19.6	**58.3**	18.2	5
León	7	10	20	2.0	1.4	2.9	**70.0**	0.0	0.0	2
Lleida	12	70	106	1.5	5.8	8.8	24.3	43.7	25.0	6
Málaga	**92**	**675**	**2,241**	3.3	7.3	24.4	39.3	24.6	53.8	**20**
Miguel Hdez	7	39	57	1.5	5.6	8.1	23.1	33.3	3.0	4
Murcia	53	274	677	2.5	5.2	12.8	40.1	38.1	28.4	12
UNED	26	171	434	2.5	6.6	16.7	35.7	28.6	17.8	10
Oviedo	51	181	527	2.9	3.5	10.3	43.1	33.3	39.0	11
Pablo Olavide	3	52	135	2.6	**17.3**	**45.0**	32.7	17.6	18.4	6
País Vasco	69	324	1,311	4.0	4.7	19.0	39.2	34.9	32.9	17
Polit. Cartagena	8	30	29	1.0	3.8	3.6	36.7	27.3	18.2	3
Polit. Catalunya	**143**	**1,175**	**4,169**	3.5	8.2	29.2	41.5	34.6	**60.5**	**29**
Polit. Madrid	**177**	**772**	**2,664**	3.5	4.4	15.1	38.2	34.4	39.3	**23**
Polit. Valencia	**175**	**1,081**	**3,226**	3.0	6.2	18.4	28.7	30.0	56.5	**25**
Pompeu Fabra	3	18	103	**5.7**	6.0	**34.3**	**94.4**	30.0	**83.3**	5
Pública Navarra	15	114	700	**6.1**	7.6	**46.7**	**64.0**	26.7	23.9	15
Rey Juan Carlos	20	186	322	1.7	9.3	16.1	37.1	27.4	23.9	10
Rovira i Virgili	23	192	693	3.6	8.3	30.1	37.5	44.3	30.8	13
Salamanca	25	138	289	2.1	5.5	11.6	31.2	19.4	25.9	8
S. Compostela	29	210	802	3.8	7.2	27.7	50.9	40.6	19.6	13
Sevilla	41	283	651	2.3	6.9	15.9	30.7	18.2	58.5	13
València	35	115	418	3.6	3.3	11.9	53.0	**45.4**	37.5	11
Valladolid	32	108	172	1.6	3.4	5.4	28.7	40.7	40.4	7
Vigo	20	50	60	1.2	2.5	3.0	34.0	42.9	3.4	4
Zaragoza	38	165	508	3.1	4.3	13.4	36.4	37.0	42.4	11

- The universities with the highest number of academic staff are: Politécnica de Madrid (177), Politécnica de Valencia (175), Politécnica de Catalunya (143), Granada (93), and Málaga (92). These values influence the fact that these universities also have the highest number of publications: Politécnica de Catalunya (1,175), Granada (1,107), Politécnica de Valencia (1081), Politécnica de Madrid (772), and Málaga (675). Although the top universities are the same in both rankings, the order is different.

- Regarding citations count, Granada (9,882), Politécnica de Catalunya (4,169), Politécnica de Valencia (3,226), Politécnica de Madrid (2,664) and Málaga (2,241) rank the five top positions. This ranking changes when the ratio between citations and publications is calculated: Jaén (9.5), Granada (8.9), Pública de Navarra (6.1), Pompeu Fabra (5.7) and Girona (5.1).

- The universities with the highest number of publications per academic are: Pablo de Olavide (17.3), Granada (11.9), Carlos III (10.7), Complutense (9.4), and Autónoma de Madrid (9.4). In the same context, the ranking of universities with the highest number of citations per academic are: Granada (106.3), Jaén (54.9), Pública de Navarra (46.7), Pablo de Olavide (45.0) and Pompeu Fabra (34.3).

- Analyzing the type of publication, the universities have different behaviors. Universities that usually publish in journals are: Pompeu Fabra (94.4%), León (70.0%), Córdoba (66.3%), Pública de Navarra (64.0%) and Granada (59.5%). In contrast, universities that usually publish in conferences are: Las Palmas de Gran Canarias (19.6%), Miguel Hernandez (23.1%), La Laguna (24.1%) and Lleida (24.3%).

- The universities with the highest percentage of documents published by Q1 journals are: Córdoba (67.3%), Las Palmas de Gran Canarias (58.3%), Burgos (57.1%), Barcelona (52.6%) and València (45.4%).

- Regarding international collaboration, the top positions in this ranking are: Pompeu Fabra (83.3%), Girona (79.3%), Cantabria (64.3%), Jaume I (61.1%) and Politécnica de Catalunya (60.5%).

- Regarding the h-index value, the universities with the highest values are: Granada (50), Politécnica de Catalunya (29), Politécnica de Valencia (25), Politécnica de Madrid (23) and Málaga (20).

8.2.4 Academic staff results

Table 8.7 analyzes the scientific research of Spanish academic staff by subject areas (CAT, CSAI, CLS) and professional standing (FP, AP1, AP2, AP3). Analyzing the number of academics belonging to each area, CLS area has the highest value. This area has also the highest number of publications, whereas the CSAI area receives the highest number of citations. Furthermore, CSAI academics also receives 4.5 citations per publication, whereas CLS

Table 8.7: Scientific production of Spanish academic staff by subject areas and professional standing.

Area	A	P	C	P/C	P/A	C/A	JRN%	Q1%	COL%	H
CAT	570	3,151	7,165	2.3	5.5	12.6	31.2	30.0	12.8	30
CSAI	586	4,222	19,181	4.5	7.2	32.7	47.0	33.8	13.2	56
CLS	848	5,049	14,744	2.9	6.0	17.4	33.3	26.5	15.3	43
Position	A	P	C	P/C	P/A	C/A	JRN%	Q1%	COL%	H
FP	280	5,722	26,213	4.6	20.4	93.6	42.6	30.6	15.3	61
AP1	1,182	8,541	23,651	2.8	7.2	20.0	35.1	30.4	12.0	50
AP2	56	326	914	2.8	5.8	16.3	44.8	30.6	7.1	14
AP3	486	638	850	1.3	1.3	1.7	24.8	29.4	5.8	12

and CAT academics receive 2.9 and 2.3 citations per publications, respectively. Regarding the number of publications and citations per academic, researchers belonging to the CSAI area have the highest values: 7.2 documents per academic and 32.7 citations per academic. By document type, CAT academics have a higher percentage of documents published in conferences than other academics. In contrast, CSAI academics have the highest percentage of documents published in journals. These academics have also the highest percentage of documents published in Q1 journals. Regarding international collaboration, CLS academics publish a 15.3% of their documents with international institutions, whereas CSAI and CAT academics achieve 13.2% and 12.8%, respectively. Finally, the *h-index* values of CSAI, CLS and CAT academics are 56, 43 and 30, respectively. On the other hand, Table 8.7 also shows that most of academics are associated with the AP1 position. The AP1 group of academics have the highest number of publications, whereas the group of FP academics receive the highest number of citations. These FP academics have also the highest values for citations per publication, publications per academics, citations per academic, percentage of international collaborations and h-index value. Analyzing the type of publications, AP3 academics have the highest percentage of documents published in conferences, whereas AP2 academics have the highest percentage of documents published in Q1 journals. Finally, Figure 8.4 reflects all academic staff over the x-axis (number of publications) and the y-axis (number of citations). Analyzing this figure, it is observed that the top-performing researchers are FP academics. These academics are associated with the three subject areas but CSAI academics usually predominate in top positions.

8.3 Discussion and conclusions

Despite its limitations, bibliometric analysis are an increasingly important topic for the scientific community. This bibliometric analysis provides a comprehensive overview of the current situation of the Spanish computer sciences area. It could be considered as a tool, which could also help decision-makers in the processes of strategic planning, in verifying the effectiveness of policies and initiatives for continuous improvement, in the optimization of limited economic

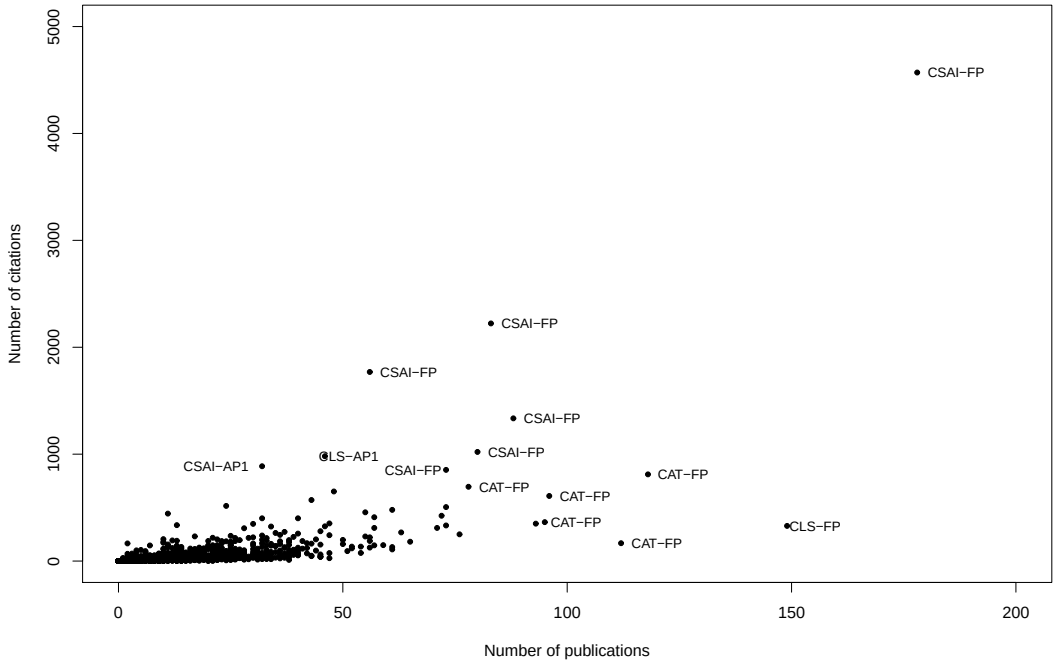


Figure 8.4: Spanish academic staff reflected over productivity and visibility axes.

resources, and in the promotion of academic staff, among others.

This overview shows that Spanish research productivity and visibility have increased their values in the last years, achieving an increment of 347% and 1,053%, respectively. Results show that Spanish academics usually publish more proceeding papers than journal articles despite of the low number of citations received by proceeding papers. Their documents are usually published in Theory and Methods and Artificial Intelligence categories. Nowadays, academics publish more documents in high quality journals than previous years. Also, Spanish academics now collaborate more with international institutions. Regarding universities, they have different behaviors in terms of disciplines, document types, collaboration types, etc. Despite this, some universities such as Universidad de Granada, Universidad Politécnica de Catalunya, Universidad Politécnica de Valencia, Universidad Politécnica de Madrid and Universidad de Málaga usually rank top positions for different parameters. By subject areas, CLS academics publish the highest number of Spanish computer science documents, whereas CSAI academics excel in terms of citation per document, documents per academic, citations per academic and percentage of documents published in high quality journals. Finally, FP and CSAI academics usually overcome other academics in most of the analyzed parameters.

According to our nationwide results, policy-makers should advise academics to increase their number of citations per document. As stated in the introduction, Spain considerably drops in the ranking when citations per document is analyzed. The objective is to improve the above ratio instead of the number of total publications. According to the document types, the

number of documents published by journals should also be improved. It is suggested because Spanish academics are assessed according to ANECA criterions which rate journal articles better than proceeding papers. Regarding collaboration, only 13.9% of documents have been published with international collaboration. This percentage has to be increased since it is well-know that publications with international collaborations usually receive more citations than other publications. Finally, another important aspect to improve is the productivity and visibility of specific areas like Cybernetics, Interdisciplinary Applications and Hardware and Architecture which have a medium-low number of publications and citations. Analyzing Spanish public universities, some universities, like Universidad de Cádiz and Universidad de Las Palmas de Gran Canaria, among others, should increase their productivity because they have few publications per academic. Similarly, universities, like Universidad Politécnica de Cartagena and Universidad de Vigo, should improve their ratio of citations per document, whereas Universidad de La Rioja, Universidad de La Laguna should publish in journals with high impact factors. By collaboration types, academics belonging to Universidad de Huelva and Universidad de León should increase their percentages of documents published with international collaboration. Finally, results could vary depending on the database consulted, which is a point to be taken into account.

Chapter 9

Cluster Analysis in Scientometrics

9.1 Introduction

The process of evaluation of scientific research has become a central element in the management and governance policies of national research systems. The most widespread evaluation methodologies can be classified into two general types: peer-review and bibliometric techniques. Although peer review is assumed to be the most reliable methodology, it is slow, expensive and unwieldy. Other authors contest this appraisal. This difference of opinion among authors has led to the development of methodologies based on bibliometric techniques. Both types of methodologies have pros and cons, extensively discussed in terms of costs, execution times, limitations and objectiveness of measurement.

Some methodologies have been published in the literature for assessing the research performance at different levels. First, Abramo and D'Angelo [5] suggested a bibliometric methodology for large-scale comparative evaluation of research performance by individual researchers, research groups and departments within research institutions. Second, Abramo *at al.* [6] also developed a bibliometric-non-parametric methodology for measuring the performance of research activities in the university system. Third, Costas *et al.* [102] proposed a general bibliometric methodology for informing the assessment of research performance of individual scientists. Finally, Torres-Salinas *et al.* [439] proposed a methodology for comparing academic institutions.

These above methodologies just presented absolute values for bibliometric indicators achieved for individual scientists, research groups, departments and institutions, among others. Other studies just performed simple descriptive exercises. In this way, Rojo and Gómez [381] provided an overview of scientific (publications) and technological (patents) production, whereas Torres-Salinas *et al.* [440] analyzed Spanish universities according to quantitative and qualitative measures related to production, impact and journal quality. Unlike previous works, Palomares-Montero and García-Aracil [352] performed a fuzzy clustering algorithm to analyze Spanish universities. They grouped universities according to three aspects (teacher mission, research mission and knowledge transfer mission) which included indicators such

as the student-teacher ratio, thesis supervised by professor, patent-teacher ratio, contracts-teacher ratio, and grants income by fulltime teacher, among others.

The objective of this chapter is to develop a cluster analysis methodology for measuring the performance of research activities in terms of productivity, visibility, quality, prestige and internationalization, while overcoming some of the limitations related to works that have been proposed in the literature. The proposed cluster analysis methodology is based on bibliometric techniques and, therefore, has many advantages (objectivity, rapidity, and low costs, among others) over a peer-review methodology. This methodology does not depend on the quality judgment of experts, so it does not suffer severe limitations related to subjectivity. It also overcomes the traditional limits of bibliometric analyses based on simple rankings and permits a robust multi-dimensional cluster analysis at the level of universities and academic staff. The cluster analysis methodology has been applied to the Spanish public universities and their academic staff in the computer science area. The results can be used to characterize the research activity of universities and academic staff, identifying both their strengths and weaknesses. These analyses afford a comprehensive overview of the current situation in the area of computer sciences in Spain.

Using the proposed methodology, policy-makers could discover knowledge related to universities and their staff. The goal of the cluster analysis methodology is to form different clusters, maximizing within-cluster homogeneity and between-cluster heterogeneity. In this way, universities/academics that belong to the same cluster are very similar to each other, whereas universities/academics belonging to different clusters are very different in term of bibliometric data. Each cluster is interpreted as providing a characterization of research activity by universities and academic staff, identifying both their strengths and weaknesses. These value-added clusters could have potential implications on research policy. Finally, this methodology supports institutions in the processes of strategic planning, in verifying the effectiveness of policies and initiatives for continuous improvement. This work appears in the published paper [226].

Chapter outline

Section 9.2 describes the procedures on which the cluster analysis methodology is based. Section 9.3 presents how both Spanish universities and their academic staff are grouped into different clusters. Finally, Section 9.4 contains some discussions and conclusions about the results and future research on the topic.

9.2 Cluster Analysis Methodology

The proposed cluster analysis methodology is divided into several procedures. These procedures are in charge of defining and describing the bibliometric variables, collecting bibliometric records from different databases, ensuring the reliability of data, calculating bibliometric indices, presenting statistical description of bibliometric indices, performing partitional, hi-

erarchical and probabilistic cluster analysis at different levels, visualizing clustering results, identifying the achieved clusters and, finally, supporting institutions on research policy decisions. These procedures are detailed in the following sections.

9.2.1 Definition of bibliometric variables

The bibliometric indices used in this methodology are widely accepted among the scientific community, measure different aspects of scientific activities and are easily interpretable. The selected bibliometric indices are detailed as follows:

- *Normalized documents*: This measure indicates the ability of each university to produce scientific knowledge. *Normalized documents* is defined as the ratio between the number of documents published by each university and the number of academics affiliated with that university. It is calculated allowing for the influence of university size in order to obtain a fair measure of production.

- *Normalized citations*: This indicator shows the scientific impact that each university has on the scientific community. It again allows for the influence of university size. *Normalized citations* is the ratio between the number of citations received by each university and the number of academics affiliated with that university.

- *Journal publication*: This measure analyzes the penchant towards either of the two most important types of research output (journals or conferences). *Journal publication* represents the ratio between the number of documents published in journals and the total number of documents published both in journals and in conferences. This indicator establishes each university's main dissemination channel.

- *First-quartile documents*: It shows the percentage of publications that a university publishes in the world's most influential scholarly journals. Journals considered for this indicator are ranked in the first quartile of their categories as ordered by Journal Citation Reports. *First-quartile documents* is the percentage of documents published in first-quartile journals with respect to the sum published in all other quartiles.

- *Fourth-quartile documents*: This is a similar indicator to *First-quartile documents* but for the least influential scholarly journals according to *JCR* (fourth-quartile).

- *Citations per journal article*: This measure is associated with the impact of journal articles. *Citations per journal article* represents the mean number of citations received by documents published in journals. This indicator reflects the quality of journal articles published by each university.

- *Citations per proceeding paper*: This indicator is associated with the impact of proceeding papers. *Citations per proceeding paper* represents the mean number of citations received by documents published in conference proceedings. This indicator reflects the quality of proceeding papers published by each university.

- *International collaboration*: This indicator shows the ability of each university to create international research links through publications. *International collaboration* represents the percentage of publications that a university publishes in collaboration with overseas institutions.

Another two bibliometric indices (*Total documents* and *Total citations*) are used in this methodology. These indices replace, respectively, *Normalized documents* and *Normalized citations* for clusterings of academic staff

- *Total documents*: This measure indicates the ability of each academic to produce scientific knowledge. *Total documents* is defined as the number of documents published by each academic. It represents the academic's productivity.

- *Total citations*: This measure shows the scientific impact that each academic has on the scientific community. *Total citations* is defined as the number of citations received by each academic. It represents the academic's visibility.

9.2.2 Data collection

Two datasets are built to analyze the research activity of Spanish public universities and their academic staff in the computer science area. The first dataset includes the values of the above bibliometric indices of Spanish universities from the date of the first publication by an active member of their academic staff (January 1, 1973) to December 31, 2009. This dataset is used to group universities into different clusters. The second dataset includes the values of the bibliometric indices for each academic from his/her first publication until January 1, 2010. This dataset is used to group academics into different clusters.

9.2.3 Statistical description of bibliometric indices

Before performing any clustering approach, a statistical summary of bibliometric indices values are presented. The objective is to provide an overview of the performance of Spanish computer science research in terms of productivity, visibility, quality, prestige and internationalization. After computing all the eight bibliometric indices for all 48 universities and 2004 academics, box plots are shown, representing the smallest observation (extreme of the lower whisker), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation (extreme of the upper whisker) for each bibliometric index. Box plots may also indicate which observations, if any, might be considered outliers. This statistical description also shows the top five universities ranked according to our bibliometric variables. These rankings are useful to compare different universities in a one-dimensional basis.

9.2.4 Cluster analysis at different levels

This procedure is concerned with finding a structure in a collection of unlabeled elements that are characterized by several variables. The goal is to group elements in this collection

so that elements that belong to a cluster are very similar to each other, whereas different clusters are highly heterogeneous.

Different starting points and criteria usually lead to different taxonomies of clustering algorithms [143, 234, 481]. A simple agreed frame is to classify clustering techniques as partitional clustering, hierarchical clustering and probabilistic clustering, based on the properties of clusters generated. Partitional clustering groups elements exclusively, so that any element belonging to one specific cluster cannot be a member of another cluster. On the other hand, hierarchical clustering produces a hierarchical structure of clusters. Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones (agglomerative clustering) or by splitting larger clusters (divisive clustering). Finally, probabilistic clustering provides a cluster membership probability for each element, where elements have a specific probability of being members of several clusters.

One of the most important issues in cluster analysis is the evaluation of clustering results [194]. Clustering validation is concerned with determining the optimal number of clusters (the best for the input dataset) and checking the quality of clustering results. Both internal and external validity indices have been used in order to evaluate the clustering results. Internal validity indices do not require a priori information from dataset, they are based on the information intrinsic to the dataset alone, whereas external validity indices require previous knowledge about dataset.

Partitional clustering Partitioning Around Medoids (PAM) [249] was used as a representative algorithm of partitional clustering in the proposed cluster analysis methodology. It assigns a set of universities/professors into k clusters with no hierarchical structure. PAM has several advantages with regard other partitional algorithms. First, this algorithm presents no limitations on attributes types because it utilizes real data points (medoids) as the cluster prototypes (medoids do not need any computation and always exist). Second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore, it is lesser sensitive to the presence of outliers. Finally, the resulting clustering is independent of the initial choice of medoids. The final objective of this algorithm is to determine a representative university/professor (medoid) for each cluster.

Hierarchical clustering The Ward's algorithm [471] was used as an advanced hierarchical clustering procedure in the proposed cluster analysis methodology. It builds a tree of clusters called dendrogram which allows exploring data on different levels of granularity. Given k clusters, the objective of this algorithm is to reduce the k clusters to k-1 mutually exclusive clusters by considering the union of all possible pairs. Unlike other hierarchical clustering algorithms which use simple measures, it uses an analysis of variance approach to evaluate the distances between clusters. Finally, the Ward's algorithm selects the union of clusters which minimizes the heterogeneity among cluster elements. The complete hierarchical structure can be obtained by repeating this process until only one cluster remains.

Probabilistic clustering The EM algorithm [124] was used as a representative algorithm of probabilistic clustering in the proposed cluster analysis methodology. The objective of this algorithm is to find the most likely set of clusters given the data. Given a number of clusters k, this algorithm models data as a finite mixture of k probability density functions. In this way, each cluster is represented by one component of the mixture. The EM algorithm was used to find the maximum likelihood estimates of the mixing coefficient and the parameters of each distribution. Finally, all variables are modeled as conditionally independent Gaussian distributions given the cluster value. Thus, each distribution is characterized by two parameters for each variable: the mean and the standard deviation.

9.2.5 Visualization of clustering results

Several figures are presented to represent different aspects of the clustering results. After performing the partitional clustering, several tables show all universities/academics grouped into disjoint clusters and the medoid bibliometric values within each cluster. Even if universities/academics belong to the same cluster, they may behave differently depending on the bibliometric indices. In this way, several figures are presented showing cluster projection for some specific bibliometric indices. Then, the hierarchical structure of clusters (dendrogram) obtained by merging smaller clusters into larger ones is represented. This dendrogram shows how the clusters are related. By cutting the dendrogram at a target level, all universities/academics are grouped into disjoint clusters. Regarding probabilistic clustering, the mean and standard deviation values for each bibliometric variable within the resulting clusters are presented. After that, each university/academic's probability of being a member of each cluster is listed.

The resulting clusters are also visually inspected using a representation in a lower dimensional space. The goal is to obtain a three-dimensional representation that approximates our eight-dimensional bibliometric variables and check whether or not the clusters were visually distinguishable. Principal component analysis was used for this purpose. Finally, it is also plotted the distribution of academics grouped in each cluster for analyzing each cluster by areas and positions associated with each academic.

9.2.6 Identification of final clusters

Each cluster can be defined according to different research activity aspects e.g. productivity (documents per academic), visibility (citations per academic), quality (citations per journal articles and proceeding papers), prestige (first-quartile journals), and internationalization (international collaboration). Global labels (high, medium-high, medium-low and low) are set for the values of each bibliometric index in the different clustering algorithms. In this way, each cluster can be represented as a set of global labels associated with research activity aspects. Using the resulting clusters and the above labels, it could be concluded that some universities/academics produce more scientific knowledge and have a bigger scientific impact than other universities/academics, whereas other universities/academics usually publish in

the most influential journals, and thus they have a selective strategy, and finally, it could be also concluded that specific universities/academics have an excellent ability to create international research collaborations.

9.2.7 Implications on research policy

Methodologies for the evaluation of research activities have been raising an increasing amount of interest in the last few years. The conclusions of these methodologies have great relevance for the design of policies to promote research and development.

Thanks to the proposed cluster analysis methodology a comprehensive overview of the current situation in a specific discipline and region is achieved. This overview could help policy-makers for making decisions. The proposed methodology could be considered as a tool, which could also help university presidents and heads of departments and research groups in the processes of strategic planning, in verifying the effectiveness of policies and initiatives for continuous improvement, in the optimization of limited economic resources, and in the promotion of academic staff, among others. The resulting clusters are interpreted as providing characterizations of research activity by universities and academic staff, identifying both their strengths and weaknesses. Using this methodology, policy-makers could propose collaborations and alliances among universities. These universities could perhaps merge strategically in order to exploit their resources, enhance their reputation and visibility, and compete with the most active international universities.

9.3 Exploring Spanish computer science research

9.3.1 Spanish public universities

All the bibliometric indices for all 48 universities are calculated. Figure 9.1 shows the box plots of the distribution of each bibliometric index. Taking *Normalized documents* as an example, it is found that 1.4 was the value of the lower whisker, whereas 11.9 was the value of the upper whisker. The 25th percentile (Q1), 50th percentile (Q2) and 75th percentile (Q3) were 3.6, 5.8, and 7.5 documents per academic, respectively. It is also found an outlier (17.3) which corresponded with Universidad Pablo de Olavide de Sevilla. Taking another example (*Journal publication*), it is observed three outliers corresponding with Universidad Pompeu Fabra (94.4), Universidad de León (70.0) and Universidad de Córdoba (66.3). In this case, the five-number summaries were: lower whisker (19.6), 25th percentile (30.8), 50th percentile (36.5), 75th percentile (44.1), and upper whisker (64.0). Finally, it is also reported the minimum and maximum value of the analyzed indices: *Normalized documents* [1.4, 17.3], *Normalized citations* [1.3, 106.3], *Journal publication* [19.6, 94.4], *First-quartile documents* [0.0, 67.3], *Fourth-quartile documents* [0.0, 57.1], *Citations per journal article* [0.7, 18.3], *Citations per proceeding paper* [0.0, 2.0], and *International collaboration* [0.0, 83.3].

Table 9.1 shows the top five universities ranked according to the selected eight variables. Analyzing the values for all universities, it is found that Universidad de Granada (UGR)

had the highest value of *Normalized citations*. This means that the mean citations received by each academic affiliated with Universidad de Granada was 106.3 citations. It is also found that the best university according to *First-quartile documents* was Universidad de

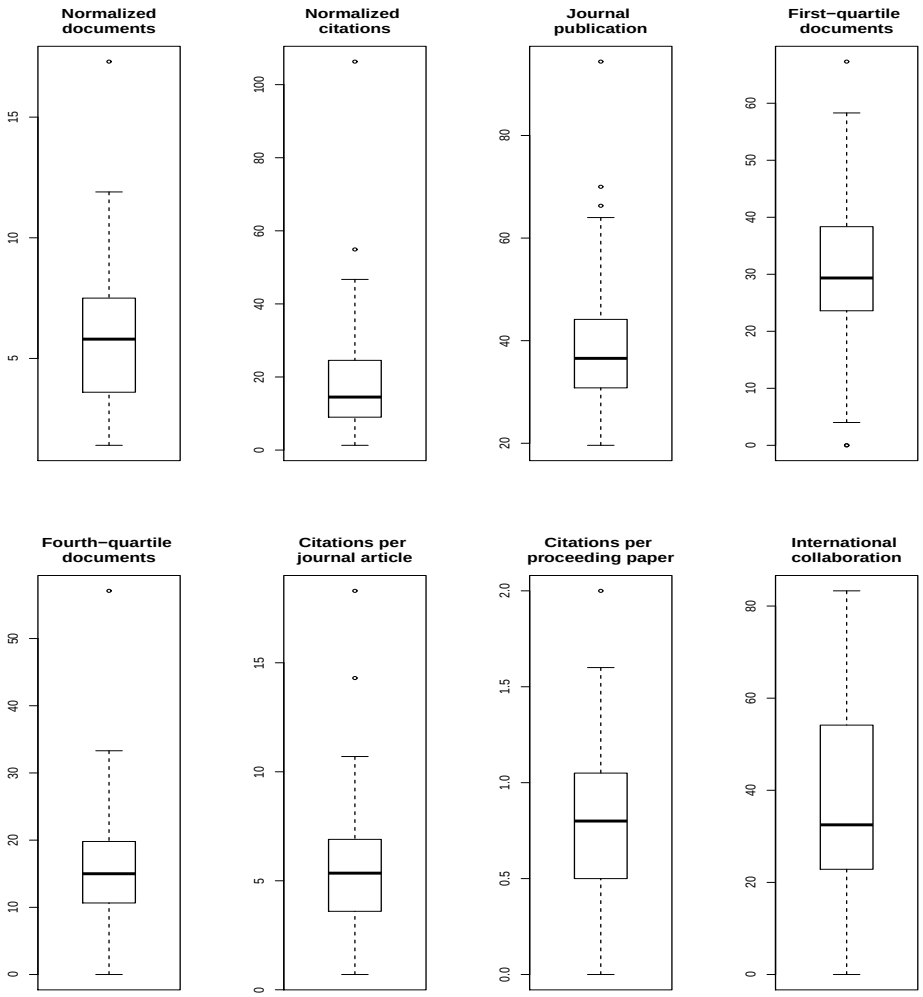


Figure 9.1: Box plot of each bibliometric index

Table 9.1: Top five universities ranked according to selected bibliometric variables

Variables	Univ 1	Univ 2	Univ 3	Univ 4	Univ 5
Normalized documents	UPO (17.3)	UGR (11.9)	UC3M (10.7)	UCM (9.4)	UAM (9.4)
Normalized citations	UGR (106.3)	UJA (54.9)	UPNA (46.7)	UPO (45.0)	UPF (34.3)
Journal publication	UPF (94.4)	ULE (70.0)	UCO (66.3)	UPNA (64.0)	UGR (59.5)
First-quartile documents	UCO (67.3)	ULPGC (58.3)	UBU (57.1)	UB (52.6)	UV (45.5)
Fourth-quartile documents	UCA (57.1)	UNEX (33.3)	ULE (33.3)	UPO (29.4)	UAB (26.2)
Citations per journal article	UJA (18.3)	UGR (14.3)	UDG (10.7)	UBU (9.6)	UPNA (9.5)
Citations per proceeding paper	UAM (2.0)	UCM (1.6)	US (1.4)	URV (1.4)	UPC (1.3)
International collaboration	UPF (83.3)	UDG (79.3)	UC (64.3)	UJI (61.1)	UPC (60.5)

Table 9.2: Partitional clustering: four clusters of universities

Clusters	Universities
Cluster A	A Coruña, Almería, Cádiz, Carlos III de Madrid, Extremadura, Huelva, La Laguna, Las Palmas de Gran Canaria, León, Lleida, Miguel Hernández de Elche, Salamanca Nacional de Educación a Distancia, Politécnica de Cartagena, Rey Juan Carlos, Vigo
Cluster B	Alcalá de Henares, Alicante, Autónoma de Barcelona, Autónoma de Madrid, Cantabria, Castilla-La Mancha, Complutense de Madrid, Girona, Jaume I de Castello, La Rioja, Málaga, Politècnica de Catalunya, Politècnica de València, Pompeu Fabra, Sevilla
Cluster C	Barcelona, Burgos, Córdoba, Illes Balears, Murcia, Oviedo, País Vasco, Valencia, Politécnica de Madrid, Rovira i Virgili, Santiago de Compostela, Valladolid, Zaragoza
Cluster D	Granada, Jaén, Pablo de Olavide, Pública de Navarra

Córdoba (UCO), that is, 67.3% of its journal articles were published in first-quartile journals. Similarly, Universidad Pompeu Fabra (UPF) was the best university regarding *International collaboration* because 83.3% of its collaborative documents were co-authored by researchers with overseas affiliations.

Before running any algorithm, the number of clusters should be fixed using clustering validation. The optimal number of clusters is usually determined based on internal validity indices like the silhouette coefficient [385]. This index is used to measure the goodness of a clustering structure without external information. Its value ranges from -1 to 1. A larger average silhouette coefficient indicates a better overall quality of the clustering result, so the optimal number of clusters is the one that gives the largest average silhouette value. After running clustering validation, it is found that the partitions with two clusters and four clusters had the highest silhouette coefficients. Although four-cluster partition had a little lower silhouette coefficient (0.65) than two-cluster partition (0.67), four-cluster partition (k=4) is selected because it more realistically explained the dataset.

After choosing the number of clusters, the partitioning around medoids (partitional clustering) is performed. Table 9.2 shows all universities grouped into four disjoint clusters. The number of universities belonging to each cluster were: cluster A (16 universities), cluster B (15 universities), cluster C (13 universities) and cluster D (4 universities).

Table 9.3 shows the medoid values within the four clusters (A, B, C and D). Analyzing the variable values, there were some differences among clusters. For example, universities belonging to cluster D had the highest value for *Normalized citations* (54.9 citations per academic). They also excelled in terms of *Journal publication* and *Citations per journal article*. Universities associated with the other clusters (A, B and C) excelled with respect to the other variables: cluster A (*Fourth-quartile documents*), cluster B (*Normalized documents, Citations per proceeding paper and International collaboration*), and cluster C (*First-quartile documents*). Finally, it shows the medoid university within each cluster. In this way, Universidad de A Coruña (UDC) was representative of cluster A, Universidad de Málaga (UMA) was representative of cluster B, Universidad Politécnica de Madrid (UPM) was representative of cluster C, and Universidad de Jaén (UJA) was representative of cluster D.

Table 9.3: Partitional clustering: Medoid values within the four clusters (A, B, C and D) and the number of universities (in parentheses) associated with each cluster

	Four resulting clusters			
Variables	A (16 univ)	B (15 univ)	C (13 univ)	D (4 univ)
Normalized documents	7.2	**7.3**	4.4	5.8
Normalized citations	10.8	24.4	15.1	**54.9**
Journal publication	29.8	39.3	38.2	**50.6**
First-quartile documents	28.2	24.6	**34.3**	24.6
Fourth-quartile documents	**17.9**	17.7	10.8	12.3
Citations per journal article	3.1	6.5	7.5	**18.3**
Citations per proceeding paper	0.8	**1.3**	1.0	0.5
International collaboration	25.2	**53.8**	39.3	21.8
Medoid university within each cluster	UDC	UMA	UPM	UJA

Even if universities belong to the same cluster, they may behave differently depending on the bibliometric indices. Figure 9.2 shows cluster analysis projection for some specific bibliometric indices. Universities belonging to clusters A, B, C and D are represented by point-down triangles, squares, circles and point-up triangles, respectively.

Figure 9.2 (top) shows the projection on the *Normalized documents* and *Normalized citations* axes. Taking cluster D (point-up triangles) as an example, there were important differences among the four universities. Universidad Pablo de Olavide (UPO) belonged to cluster D and ranked 1st for *Normalized documents*, whereas Universiad de Jaén (UJA), which also belonged to cluster D, ranked 24th for *Normalized documents*. It is also observed big differences between Universidad de Granada (UGR) and the other three universities regarding *Normalized citations*. Despite these differences, the four universities were the top scorers for *Normalized citations*.

Figure 9.2 (middle) shows the projection on the *First-quartile documents* and *Journal publication* axes. Note that these rankings are very different to the previous ones (Figure 9.2, top). According to Table 9.3, cluster A had the lowest value for *Journal publication*. Despite this, Universidad de León (ULE), which belonged to cluster A, ranked 2nd for *Journal publication*, outperforming universities belonging to better clusters. Universidad Pompeu Fabra (UPF) ranked 1st for *Journal publication* and was a member of cluster B, which was not the highest scoring group for the analyzed variable. On the other hand, universities belonging to cluster C usually ranked top for *First-quartile documents*. Universidad de Córdoba (UCO), which ranked 1st for *First-quartile documents*, also ranked 3rd for *Journal publication*.

Figure 9.2 (bottom) shows the projection on the *Citations per journal article* and *Citations per proceeding paper* axes. In this case, two universities belonging to cluster D (Universidad de Jaén (UJA) and Universidad de Granada (UGR)) were among highest scorers for *Citations per journal article*. Universities belonging to cluster B (squares) did not score high for *Citations per journal article*. Even so, Universidad de Girona (UDG), which belonged to cluster B, ranked 3rd for the above measure. On the other hand, universities belonging to cluster A (point-down triangles) did not score high on *Citations per proceeding paper*. Even

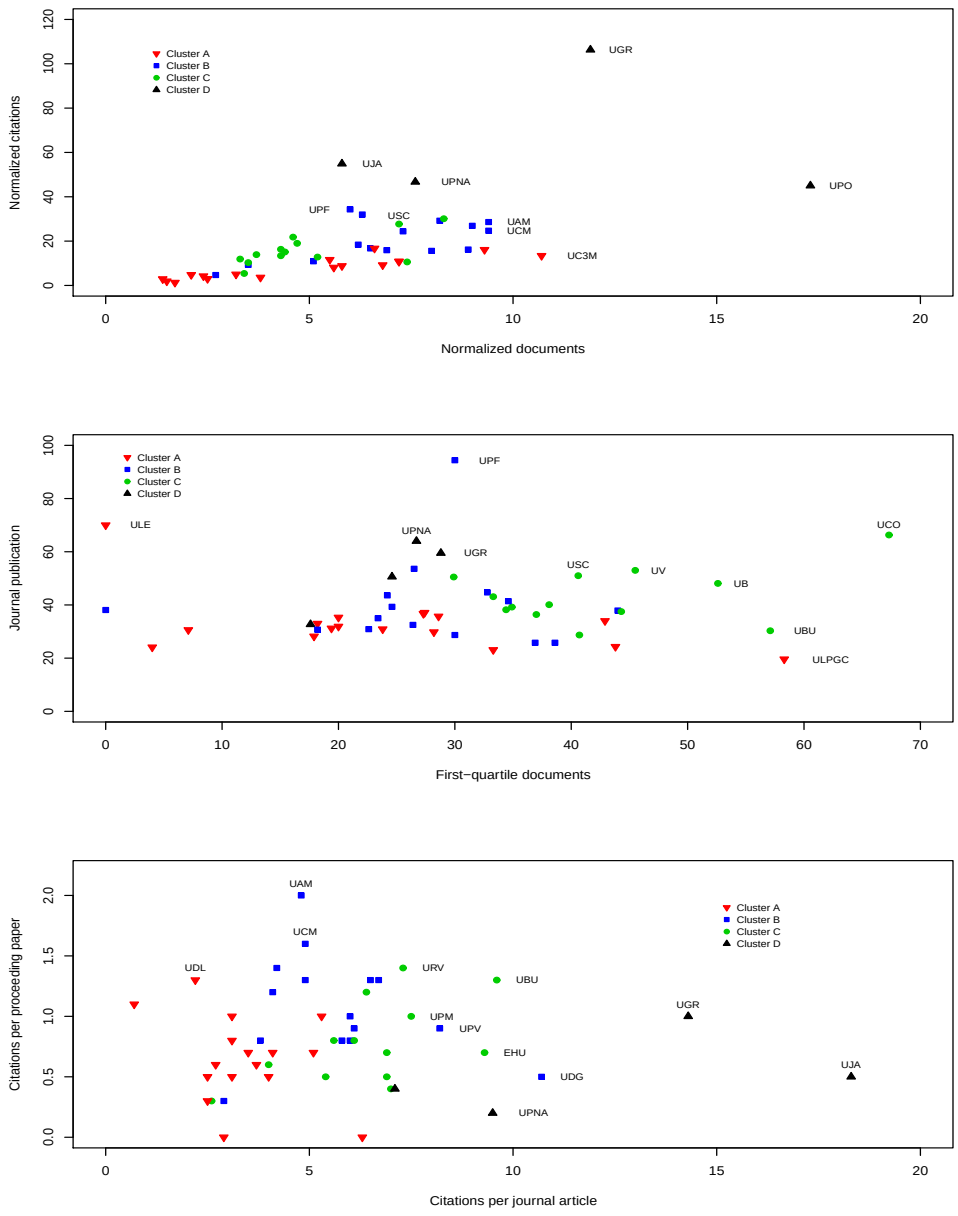


Figure 9.2: Partitional clustering: Projection on bibliometric indices axes. Universities belonging to clusters A, B, C and D are represented by point-down triangles, squares, circles and point-up triangles

so, Universidad de Lleida (UDL), which belonged to cluster A, ranked 6th for the above measure. Universities belonging to cluster B, like Universidad Autónoma de Madrid (UAM) and Universidad Complutense de Madrid (UCM), ranked top for *Citations per proceeding paper*.

Figure 9.3 represents the hierarchical structure of clusters (dendrogram) obtained by merging smaller clusters into larger ones (Ward's algorithm). This dendrogram shows how the clusters are related. By cutting the dendrogram at the horizontal line (target level),

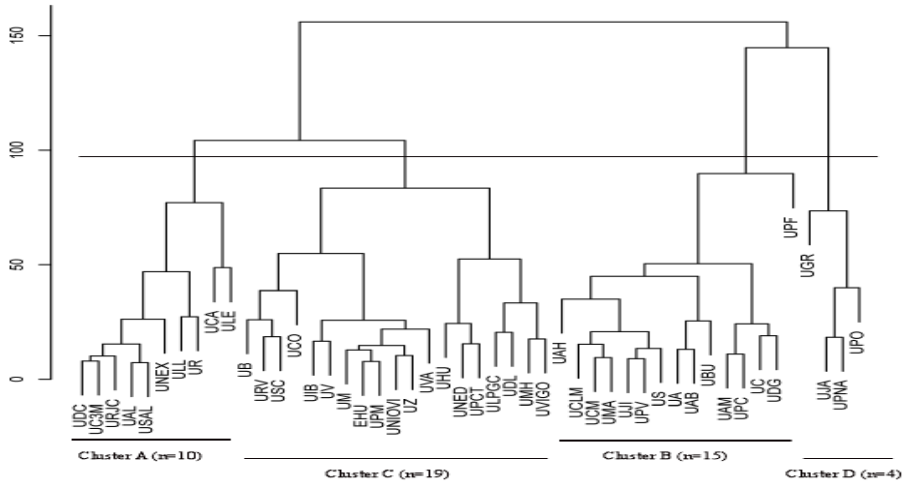


Figure 9.3: Hierarchical clustering: Hierarchical structure of clusters (dendrogram) obtained by merging smaller clusters into larger ones using Ward's algorithm

all universities are grouped into four disjoint clusters. Internal validity indices for cutting the dendrogram have been used instead of others criterions proposed by Maarek and Ben Shaul [302]. They proposed slicing techniques that automatically identify the cut-off point within the dendrogram which has a comparable degree of intra-cluster similarity.

Figure 9.3 shows that the number of universities belonging to each cluster is: cluster A (10 universities), cluster B (15 universities), cluster C (19 universities) and cluster D (4 universities). These results are very similar to the outcomes for partitional clustering. Note that the four universities belonging to cluster D, and 14 out of 15 universities belonging to cluster B are the same as before. Regarding cluster A, note that it contains fewer universities (down from 16 to 10), six universities having moved to cluster C.

Regarding probabilistic clustering, four different clusters were formed by running the EM algorithm. Table 9.4 shows the mean and standard deviation values for each variable

Table 9.4: Probabilistic clustering: Mean ± standard deviation values for each variable within the four clusters (A, B, C and D) and the number of universities (in parenthesis) associated with each cluster

	Four resulting clusters			
Variables	A (19 univ)	B (16 univ)	C (9 univ)	D (4 univ)
Normalized documents	4.6 ± 2.6	7.1 ± 1.7	4.9 ± 1.2	**10.6 ± 4.4**
Normalized citations	7.2 ± 4.1	21.9 ± 7.4	16.9 ± 5.1	**62.4 ± 25.7**
Journal publication	34.8 ± 11.4	38.5 ± 15.5	46.5 ± 9.3	**51.4 ± 12.1**
First-quartile documents	26.1 ± 16.1	32.7 ± 9.9	**39.4 ± 11.2**	24.5 ± 4.2
Fourth-quartile documents	**19.2 ± 12.4**	15.9 ± 4.6	9.6 ± 3.6	15.8 ± 7.9
Citations per journal article	3.4 ± 1.2	6.3 ± 1.8	6.7 ± 1.2	**12.2 ± 4.4**
Citations per proceeding paper	0.6 ± 0.4	**1.1 ± 0.4**	0.7 ± 0.2	0.5 ± 0.3
International collaboration	24.0 ± 15.9	**56.5 ± 12.9**	31.9 ± 7.9	23.9 ± 4.8

within the four resulting clusters. Taking *Normalized citations* as an example, universities belonging to cluster D received an average number of citations equal to 62.4±25.7 citations per academic. In contrast, universities belonging to cluster A, B, and C received on average fewer citations: 7.2±4.1, 21.9±7.4 and 16.9±5.1, respectively.

Table 9.5 shows which universities belong to each cluster. It lists each university's probability of being a member of each cluster. The highest membership probability for almost all universities was close to 1.00. This means that there was no doubt about which cluster

Table 9.5: Cluster membership probability of each university

University	Cluster A	Cluster B	Cluster C	Cluster D
A Coruña (UDC)	**0.98983**	0.00950	0.00026	0.00040
Alcalá (UAH)	**0.99902**	0.00098	0.00000	0.00000
Alicante (UA)	0.02357	**0.97587**	0.00056	0.00000
Almería (UAL)	**0.99987**	0.00003	0.00002	0.00008
Autónoma de Barcelona (UAB)	0.05224	**0.94776**	0.00000	0.00000
Autónoma de Madrid (UAM)	0.00000	**1.00000**	0.00000	0.00000
Barcelona (UB)	**0.99745**	0.00020	0.00235	0.00000
Burgos (UBU)	0.00000	**0.99723**	0.00277	0.00000
Cádiz (UCA)	**1.00000**	0.00000	0.00000	0.00000
Cantabria (UC)	0.00000	**0.99999**	0.00001	0.00000
Carlos III de Madrid (UC3M)	**0.97364**	0.01380	0.00000	0.01257
Castilla-La Mancha (UCLM)	0.19839	**0.80157**	0.00004	0.00001
Complutense de Madrid (UCM)	0.00000	**1.00000**	0.00000	0.00000
Córdoba (UCO)	0.00000	0.00003	**0.99997**	0.00000
Extremadura (UNEX)	**1.00000**	0.00000	0.00000	0.00000
Girona (UDG)	0.00000	**1.00000**	0.00000	0.00000
Granada (UGR)	0.00000	0.00000	0.00000	**1.00000**
Huelva (UHU)	**1.00000**	0.00000	0.00000	0.00000
Illes Balears (UIB)	0.00008	0.00246	**0.99746**	0.00000
Jaén (UJA)	0.00000	0.00000	0.00000	**1.00000**
Jaume I de Castellón (UJI)	0.00035	**0.99862**	0.00103	0.00000
La Laguna (ULL)	**0.99931**	0.00044	0.00025	0.00000
La Rioja (UR)	**0.94613**	0.05385	0.00002	0.00000
Las Palmas de Gran Canaria (ULPGC)	**1.00000**	0.00000	0.00000	0.00000
León (ULE)	**1.00000**	0.00000	0.00000	0.00000
Lleida (UDL)	**0.99801**	0.00199	0.00000	0.00000
Málaga (UMA)	0.00000	**0.99999**	0.00001	0.00000
Miguel Hernández de Elche (UMH)	**0.99999**	0.00001	0.00000	0.00000
Murcia (UM)	0.07037	0.00822	**0.92141**	0.00001
Nacional de Educación a Distancia (UNED)	0.09010	0.02112	**0.88724**	0.00154
Oviedo (UNIOVI)	0.08176	0.02303	**0.89521**	0.00000
Pablo Olavide de Sevilla (UPO)	0.00000	0.00000	0.00000	**1.00000**
País Vasco (EHU)	0.00000	0.00886	**0.99110**	0.00003
Politécnica de Cartagena (UPCT)	**0.99994**	0.00000	0.00000	0.00006
Politècnica de Catalunya (UPC)	0.00000	**1.00000**	0.00000	0.00000
Politécnica de Madrid (UPM)	0.00008	0.03569	**0.96423**	0.00000
Politècnica de València (UPV)	0.00000	**0.99949**	0.00051	0.00000
Pompeu Fabra (UPF)	0.00000	**1.00000**	0.00000	0.00000
Pública de Navarra (UPNA)	0.00000	0.00001	0.00000	**0.99999**
Rey Juan Carlos (URJC)	**0.82268**	0.12601	0.00047	0.05084
Rovira i Virgili (URV)	0.00000	**0.99930**	0.00070	0.00000
Salamanca (USAL)	**0.96119**	0.01620	0.02137	0.00125
Santiago de Compostela (USC)	0.00000	0.01375	**0.98619**	0.00006
Sevilla (US)	0.00394	**0.99606**	0.00000	0.00000
València (UV)	0.00117	0.00068	**0.99815**	0.00000
Valladolid (UVA)	**0.97670**	0.00070	0.02261	0.00000
Vigo (UVIGO)	**1.00000**	0.00000	0.00000	0.00000
Zaragoza (UZ)	0.01693	**0.95124**	0.03183	0.00000

they belong to. For example, the members of cluster D were: Universidad de Granada, Universidad de Jaén, Universidad Pablo de Olavide de Sevilla (their membership probability of cluster D was 1.00000) and Universidad Pública de Navarra (its membership probability of cluster D was 0.99999). On the other hand, it is observed that all universities belonging to cluster A had a probability greater than 0.82, whereas universities belonging to cluster B and C, had a probability greater than 0.80 and 0.88, respectively.

Clustering validation is also concerned with checking the quality of clustering results us-

Table 9.6: Comparisons among three different clustering results

University	Partitional	Hierarchical	Probabilistic
A Coruña (UDC)	A	A	A
Alcalá (UAH)	B	B	A
Alicante (UA)	B	B	B
Almería (UAL)	A	A	A
Autónoma de Barcelona (UAB)	B	B	B
Autónoma de Madrid (UAM)	B	B	B
Barcelona (UB)	C	C	A
Burgos (UBU)	C	B	B
Cádiz (UCA)	A	A	A
Cantabria (UC)	B	B	B
Carlos III de Madrid (UC3M)	A	A	A
Castilla-La Mancha (UCLM)	B	B	B
Complutense de Madrid (UCM)	B	B	B
Córdoba (UCO)	C	C	C
Extremadura (UNEX)	A	A	A
Girona (UDG)	B	B	B
Granada (UGR)	D	D	D
Huelva (UHU)	A	C	A
Illes Balears (UIB)	C	C	C
Jaén (UJA)	D	D	D
Jaume I de Castellón (UJI)	B	B	B
La Laguna (ULL)	A	A	A
La Rioja (UR)	B	A	A
Las Palmas de Gran Canaria (ULPGC)	A	C	A
León (ULE)	A	A	A
Lleida (UDL)	A	C	A
Málaga (UMA)	B	B	B
Miguel Hernández de Elche (UMH)	A	C	A
Murcia (UM)	C	C	C
Nacional de Educación a Distancia (UNED)	A	C	C
Oviedo (UNIOVI)	C	C	C
Pablo Olavide de Sevilla (UPO)	D	D	D
País Vasco (EHU)	C	C	C
Politécnica de Cartagena (UPCT)	A	C	A
Politècnica de Catalunya (UPC)	B	B	B
Politécnica de Madrid (UPM)	C	C	C
Politècnica de València (UPV)	B	B	B
Pompeu Fabra (UPF)	B	B	B
Pública de Navarra (UPNA)	D	D	D
Rey Juan Carlos (URJC)	A	A	A
Rovira i Virgili (URV)	C	C	B
Salamanca (USAL)	A	A	A
Santiago de Compostela (USC)	C	C	C
Sevilla (US)	B	B	B
València (UV)	C	C	C
Valladolid (UVA)	C	C	A
Vigo (UVIGO)	A	C	A
Zaragoza (UZ)	C	C	B

Table 9.7: Definition of clusters regarding different research activity aspects

	Cluster A	Cluster B	Cluster C	Cluster D
Productivity	medium-low	medium-high	medium-low	high
Visibility	low	medium-low	medium-low	high
Quality	medium-low	medium-high	medium-high	high
Prestige	medium-low	medium-high	high	medium-low
Internationalization	medium-low	high	medium-high	medium-low

ing external validity indices like the Rand index [374]. It takes the value of 1 when the two clusterings are identical. After running clustering validation, important similarities among the clustering algorithms were found: partitional vs hierarchical (0.8262), partitional vs probabilistic (0.8245), hierarchical vs probabilistic (0.7819). Note that the agreement between the hierarchical and probabilistic clustering pair had the lowest Rand index value, whereas the other clusterings had similar agreements. Table 9.6 compares the results of the three cluster algorithms to see the robustness of the results. Universities listed in grey shaded rows were not grouped in the same cluster by all the cluster algorithms.

Each cluster can be defined according to different research activity aspects (e.g. productivity (documents per academic), visibility (citations per academic), quality (citations per journal articles and proceeding papers), prestige (first-quartile journals), and internationalization (international collaboration)). Global labels (high, medium-high, medium-low and low) are set for the values of each bibliometric index in the different clustering algorithms. Table 9.7 represents each cluster according to research activity aspects. Taking cluster B as an example, the productivity of universities belonging to this cluster was medium-high but visibility was medium-low. Also, their values for quality, prestige and internationalization were medium-high, medium-high and high, respectively.

In order to summarize all results, it is concluded that universities belonging to cluster D produce more scientific knowledge and have a bigger scientific impact than other universities. Universities belonging to cluster C usually publish in the most influential journals, and thus they have a selective strategy. In contrast, universities belonging to cluster B have an excellent ability to create international research publications, whereas universities belonging to cluster A do not stand out on any research activity aspect.

Finally, other variables like the *number of computer science theses* published during the 2005-2009 period, which was not used for the clustering, is also used as a external variable to describe the four resulting clusters. Results show that cluster B had the highest value, followed by cluster D, cluster C and cluster A. Analyzing all universities, it is also observed that the three top ranked universities for *number of computer science theses* were Universidad Politècnica de Catalunya (cluster B), Universidad Politècnica de València (cluster B) and Universidad de Granada (cluster D).

Table 9.8: Mean ± standard deviation values for each variable within the six clusters (A, B, C, D, E and F) and the number of academics (in parentheses) associated with each cluster

	Six resulting clusters					
Variables	A (321)	B (839)	C (416)	D (166)	E (248)	F (14)
TD	14.0 ±9.4	2.5 ±4.8	9.2 ±8.9	33.1 ±20.4	11.0 ±8.1	**74.5 ±39.8**
TC	26.4 ±23.2	2.3 ±6.5	19.8 ±21.5	175.5 ±89.0	21.0 ±20.4	**1249.4 ±1071.7**
JP	37.7 ±24.9	2.8 ±7.6	55.9 ±26.5	54.7 ±22.1	42.1 ±24.5	**69.6 ±14.7**
Q1	23.6 ±28.6	0.0 ±0.9	5.6 ±10.8	31.7 ±19.9	**70.6 ±25.6**	37.4 ±15.0
Q4	14.2 ±23.5	0.0 ±0.0	**28.4 ±36.1**	11.8 ±13.0	3.9 ±11.1	8.9 ±7.7
CJ	4.3 ±6.4	0.4 ±1.7	4.1 ±5.0	11.0 ±8.6	4.5 ±6.3	**23.5 ±8.7**
CP	0.8 ±0.9	0.3 ±0.9	0.6 ±0.8	**1.4 ±1.6**	0.6 ±0.7	0.9 ±0.7
IC	**83.5 ±20.0**	0.6 ±4.3	5.5 ±12.3	48.0 ±31.7	8.7 ±16.5	39.8 ±19.5

TD=Total documents, *TC=Total citations*, *JP=Journal publication*, *Q1=First-quartile documents*, *Q4=Fourth-quartile documents*, *CJ=Citations per journal article*, *CP=Citations per proceeding paper*, and *IC=International collaboration*

9.3.2 Spanish public university academic staff

All bibliometric indices for all 2004 academics were calculated. It is reported the minimum and maximum value of the distribution of each selected bibliometric index: *Total documents* [0, 178], *Total citations* [0, 4570], *Journal publication* [0, 100], *First-quartile documents* [0, 100], *Fourth-quartile documents* [0, 100], *Citations per journal article* [0, 82.5], *Citations per proceeding paper* [0, 16.0], and *International collaboration* [0, 100]. Taking *Total citations* as an example, it is found that 0 was the lowest number of citations received by a specific academic, whereas 4570 was the highest value.

An internal clustering validation was also performed to find the optimal number of clusters for the academic staff dataset. After running clustering validation, the partition with six clusters (k=6) had the highest silhouette coefficient. In this way, a partitional clustering algorithm (partitioning around medoids) was run setting the number of clusters to six. Hierarchical and probabilistic clusterings were not performed for space reasons. Figures associated with these cluster analyses were very big for representing 2004 academics.

Table 9.8 shows the number of academics (in parentheses) associated with each cluster and the mean and standard deviation values for each variable within the six resulting clusters. It is observed that the number of academics belonging to each cluster were: cluster A (321 academics), cluster B (839 academics), cluster C (416 academics), cluster D (166 academics), cluster E (248 academics), and cluster F (14 academics). Analyzing the variable values, there were some differences among clusters. Taking *Total documents* as an example, it is observed that academics belonging to cluster F had the highest mean value (74.5±39.8). They also stood out on *Total citations*, *Journal publication* and *Citations per journal article*. Academics associated with cluster E excelled in terms of *First-quartile documents*, whereas academics associated with cluster C excelled with respect to *Fourth-quartile documents*. Finally, academics in cluster D had the highest value of *Citations per proceeding paper* and academics belonging to cluster A stood out on *International collaboration*.

The clusters obtained with partitioning around medoids were visually inspected using

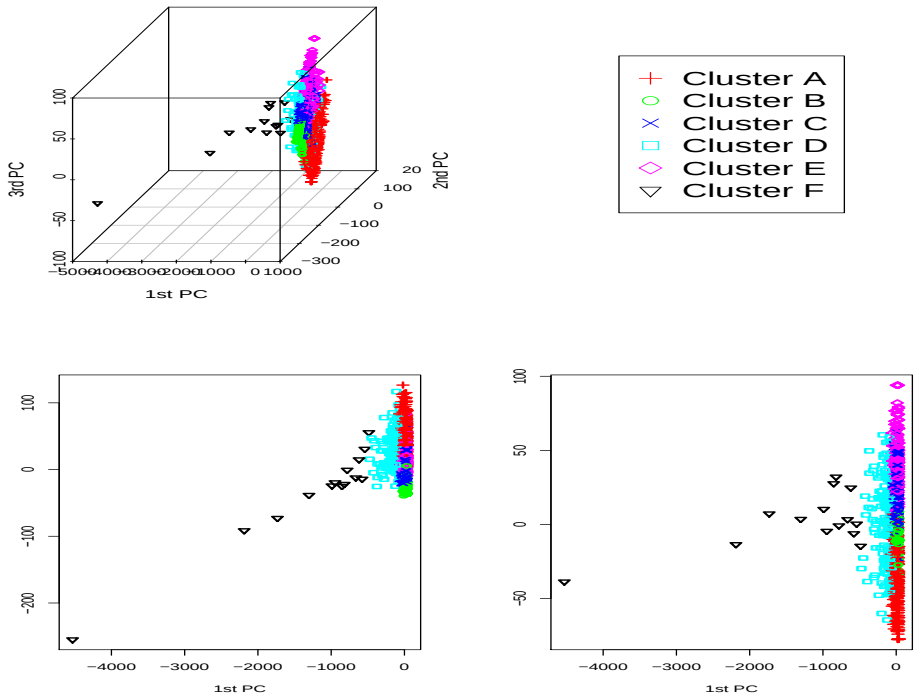


Figure 9.4: Visualization of the academic clusters in three and two-dimensional spaces obtained with principal component analysis

a representation in a lower dimensional space (see Figure 9.4). The goal was to obtain a three-dimensional representation that approximates our eight-dimensional variables and check whether or not the clusters were visually distinguishable. A principal component analysis [359] was performed, and the three principal components which account for the highest proportion of variance (95.0%) were studied.

Figure 9.4 plots the values of the bibliometric indices for each academic in the transformed three-dimensional space. Different symbols and colors were used to show the cluster assigned by the clustering algorithm to each academic. Two-dimensional projections were also included for ease of interpretation. The first principal component (1st PC), which accounted for 85.9% of the variance, distinguished academics in cluster D and cluster F from the other clusters. The second principal component (2nd PC) distinguished academics belonging to cluster A from cluster B, and accounted for 5.9% of the variance. Finally, the third principal component (3rd PC), which accounted for 3.2% of the variance, distinguished between academics belonging to cluster E and cluster A. It also distinguished between academics belonging to cluster E and cluster B.

Table 9.9 shows the number of academics at each university belonging to each of the six clusters. Taking cluster F as an example, it is observed that its 14 members were: 1 academic from Universidad de Girona, 8 academics from Universidad de Granada, 1 academic from

Table 9.9: Number of academics within the six clusters by universities

	Six resulting clusters					
Academics from university	A (321)	B (839)	C (416)	D (166)	E (248)	F (14)
A Coruña	6	16	15	0	8	0
Alcalá	6	23	9	1	2	0
Alicante	13	30	14	4	16	0
Almería	3	21	8	0	5	0
Autónoma de Barcelona	12	12	11	7	4	0
Autónoma de Madrid	10	6	5	3	3	0
Barcelona	0	3	1	0	3	0
Burgos	1	7	0	1	0	0
Cádiz	1	22	9	0	0	0
Cantabria	5	2	4	1	1	0
Carlos III de Madrid	4	2	13	1	7	0
Castilla-La Mancha	12	18	15	7	15	0
Complutense de Madrid	17	19	11	9	9	0
Córdoba	0	6	3	5	6	0
Extremadura	6	29	8	0	3	0
Girona	13	8	2	2	1	1
Granada	8	11	27	33	6	8
Huelva	0	6	2	0	0	0
Illes Balears	5	23	6	5	8	0
Jaén	2	12	7	1	4	1
Jaume I de Castellón	18	24	10	3	3	0
La Laguna	6	4	7	0	0	0
La Rioja	1	3	2	0	0	0
Las Palmas de Gran Canaria	0	45	6	0	7	0
León	0	3	4	0	0	0
Lleida	4	3	3	0	2	0
Málaga	12	30	22	13	15	0
Miguel Hernández de Elche	0	5	1	0	1	0
Murcia	7	22	15	3	6	0
Nacional de Educación a Distancia	3	9	9	2	3	0
Oviedo	7	32	7	1	4	0
Pablo Olavide de Sevilla	0	0	0	1	2	0
País Vasco	5	37	13	6	8	0
Politécnica de Cartagena	0	5	1	0	2	0
Politècnica de Catalunya	45	27	29	20	22	0
Politécnica de Madrid	19	106	30	10	11	1
Politècnica de València	41	70	32	13	17	2
Pompeu Fabra	2	0	0	0	1	0
Pública de Navarra	3	1	3	1	6	1
Rey Juan Carlos	3	3	10	0	4	0
Rovira i Virgili	0	15	5	2	1	0
Salamanca	1	20	1	1	2	0
Santiago de Compostela	1	12	8	4	4	0
Sevilla	6	18	10	3	4	0
València	3	21	6	1	4	0
Valladolid	6	19	3	0	4	0
Vigo	0	14	4	0	2	0
Zaragoza	4	15	5	2	12	0

Universidad de Jaén, 1 academic from Universidad Politécnica de Madrid, 2 academics from Universidad Politècnica de València and 1 academic from Universidad Pública de Navarra. Moreover, the biggest cluster B was composed mainly of a group of academics from the Universidad Politécnica de Madrid (106), Universidad Politècnica de València (70) and Universidad de Las Palmas de Gran Canaria (45).

Summarizing all results, academics of cluster F usually produced more scientific knowl-

edge and had more impact than other academics. They had the highest impact in terms of journal articles, whereas academics belonging to cluster D excelled with respect to proceeding papers. Academics belonging to cluster E published in the most influential journals, whereas academics belonging to cluster C published in journals with lower impact factors. Academics of cluster A stood out for their ability to author international research publications. Finally, academics belonging to cluster B did not stand out on any research activity aspect.

By areas (CAT, CSAI, CLS) and positions (FP, AP1, AP2, AP3) associated with each academic, Figure 9.5 shows the distribution of academics grouped in each cluster. For example, cluster F had 14 members, 4 of whom (28.6%) work on CAT, 8 (57.1%) on CSAI, and 2 (14.3%) on CLS. Also, cluster F was composed of 4 FP working on CAT, 6 FP and 2 AP1 working on CSAI, and 1 FP and 1 AP1 working on CLS. Figure 9.5 also shows that cluster A, cluster B, cluster C and cluster D were mainly composed of CLS academics, whereas members of clusters D and F were mainly CSAI academics. Taking into account academic positions, cluster A, cluster C, and cluster E were mainly composed of AP1, cluster B was mainly composed of AP1 and AP3, cluster D was mainly composed of FP and AP1, and finally, cluster F was mainly composed of FP.

9.4 Discussion and conclusions

This chapter proposes a cluster analysis methodology to evaluate the research activity (in terms of bibliometric indices) of institutions and their academic staff. This analysis focuses on the study of Spanish public universities and academics working in the computer science field, but this methodology can also be applied in other academic settings as well as in other research areas and countries.

The proposed methodology offers a series of advantages when it is compared to the classic peer review methodologies. Specially, it does not suffer limitations related to subjectivity since it does not depend on the quality judgment of experts. It is an objective technique for assessing research performance, overcomes the traditional limits of bibliometric analyses based on simple rankings and permits a multi-dimensional cluster analysis at different levels.

This cluster analysis methodology groups similar universities or academics in the same cluster, maximizing within-cluster homogeneity and between-cluster heterogeneity. These results are useful for characterizing the research activity of universities and their academic staff. Three well-known clustering approaches (partitional, hierarchical and probabilistic) are used to give a comprehensive overview of the current situation by means of their useful different outputs (cluster medoids, dendrograms and cluster probabilities, among others). Other clustering approaches, such as combinatorial search-based techniques, kernel-based techniques, graph theory-based techniques, neural networks-based techniques and fuzzy techniques, have not been used in this methodology. Further analysis, including the above approaches, could give a more sophisticated overview. Regarding clustering validation, the silhouette coefficient and Rand index were used to determine the optimal number of clusters and the agreement between two different partitions, respectively. Other internal indices (e.g. Dunn index) and

external indices (e.g. Adjusted Rand index), have been also used but the results (not shown) did not vary so much.

Spanish public universities were grouped into four different clusters. Universities that belong to cluster D (Universidad de Granada, Universidad de Jaén, Universidad Pablo de Olavide de Sevilla and Universidad Pública de Navarra) score highest for the following research

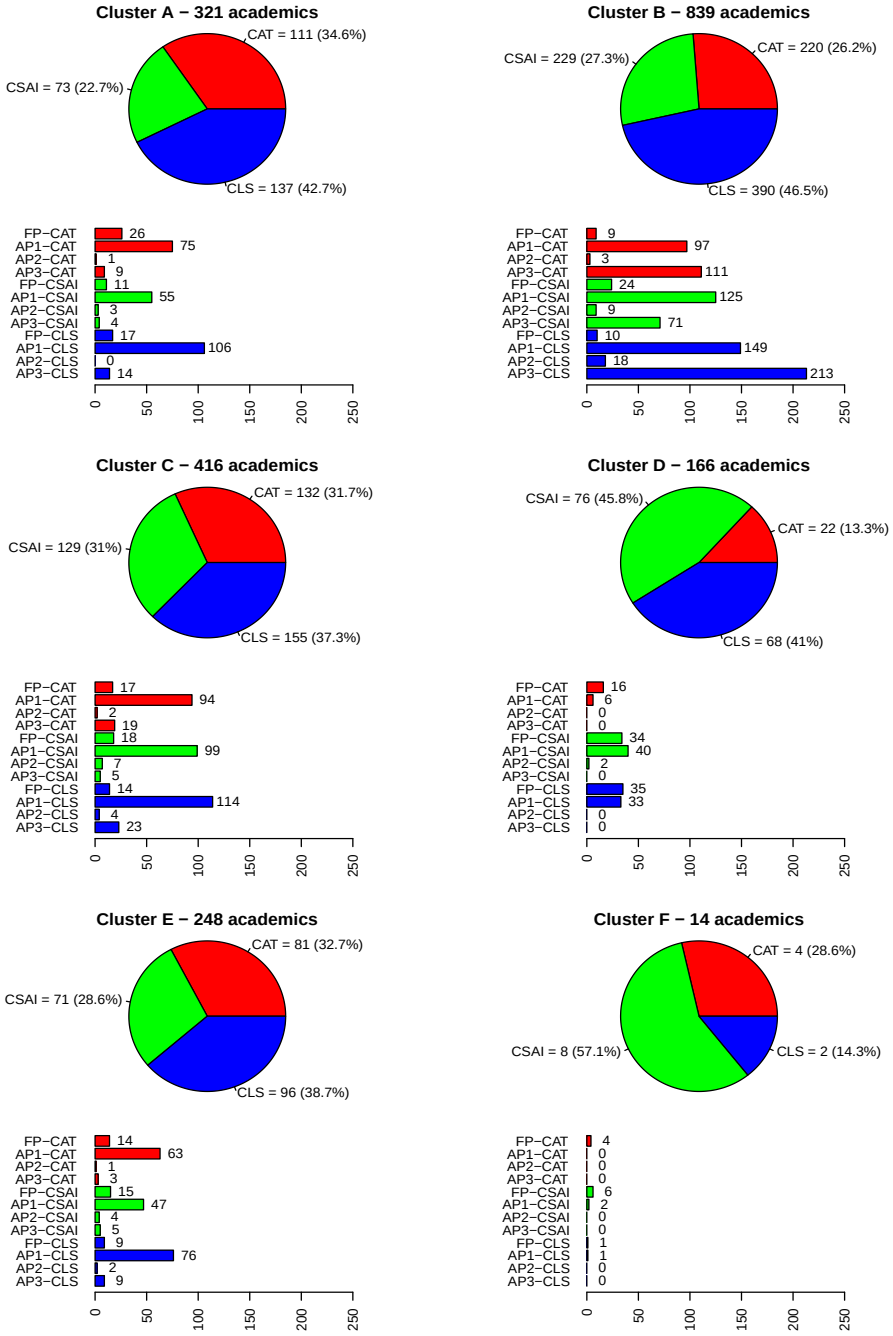


Figure 9.5: Distribution of academics belonging to each cluster by areas and positions

activity aspects: productivity, visibility and quality. Universities belonging to cluster C (Universidad de Córdoba, Universidad del País Vasco, and Universidad Politécnica de Madrid, among others) excel in terms of prestige, whereas universities belonging to cluster B (Universidad de Girona, Universidad Politécnica de Valéncia, and Universidad Pompeu Fabra, among others) stand out on international collaboration. Finally, universities belonging to cluster A have worse scores for research activity aspects than the other universities.

Unlike Bornmann and Leydesdorff [54] who showed that northern cities perform better than southern cities in some countries like Italy, it is found that most of universities belonging to cluster D, which score highest for productivity, visibility and quality, are southern universities. In contrast, some northern universities like Universidad Autónoma de Barcelona, Universidad de Cantabria, Universidad de Girona, Universidad Politécnica de Catalunya stand out on international collaboration, whereas Universidad de Oviedo, Universidad del País Vasco, and Universidad Santiago de Compostela excel in terms of prestige.

Spanish computer science output originates mainly in higher education institutions. Analyzing Spanish university results, it is noted that they do not stand out for their quality. Citations per document is used as a research quality indicator in order to compare Spanish universities with other international universities. According to Essential Science Indicators, ten Spanish universities rank in the top 350 positions, but only two (Universidad de Barcelona and Universidad de Vigo) are among the top 100 for citations per document. A possible reason for this situation is the constant cuts in the Spanish science budget [351].

The cluster analysis methodology grouped Spanish academics into six different clusters: cluster A (321 academics), cluster B (839 academics), cluster C (416 academics), cluster D (166 academics), cluster E (248 academics), and cluster F (14 academics). Each cluster can be summarized with respect to different research activity aspects. Academics belonging to cluster F excel in terms of productivity, visibility and quality, whereas academics belonging to cluster E and cluster A stand out for their prestige and internationalization, respectively. Other academics that belong to clusters B, C, and D score worse in terms of research activity aspects. Focusing on cluster F (the best in terms of productivity, visibility and quality), academics from Universidad de Girona, Universidad de Granada, Universidad de Jaén, Universidad Politécnica de Madrid, Universidad Politècnica de València and Universidad Pública de Navarra are members of cluster F. Also, this cluster is composed mainly by full professors of the computer science and artificial intelligence area.

Agrait and Poves [12] stated that not all Spanish academics publish research. Even so, they have paid time for researching. Their results show that 43.7% of Spanish academics regularly publish documents or patents. By positions, results show that 69.5% of FP, 40.6% of AP1, 21.5% of AP2 and 4.9% of AP3 usually do research. These results corroborate the findings in Spanish computer science. Cluster B includes the highest number of academics (839 out of 2004). Table 9.8 shows that the academics in cluster B have a low score for publications and citations. Also, this cluster is mainly composed of AP3 academics (see Figure 9.5).

The proposed cluster analysis methodology can help institutions to compare themselves

to each other and motivate them to improve theirs outcomes, since the methodology characterize research activity, identifying both their strengths and weaknesses. According to the results, academic researchers should improve the quality (number of citations per paper) in universities belonging to cluster A, the visibility (number of citations per academic) in universities belonging to cluster B, the productivity (number of publications per academic) in universities belonging to cluster C, and prestige (the percentage of documents published in first-quartile journals) in universities belonging to cluster D. On the other hand, academics belonging to AP2 and AP3 positions should increase their productivity and visibility, AP1 academics working in CAT and CLS should improve their quality, AP1 academics working in CSAI should publish in journals with higher impact factor, and finally, FP academics should collaborate with foreign institutions.

Using the cluster analysis methodology, policy-makers could propose collaborations and alliances among universities belonging to the same cluster. Several universities should perhaps merge strategically in order to compete with the most active international universities. In this way, Spanish universities could exploit their resources, enhance their reputation and visibility, and rise in the international rankings.

In the future, the target will be to incorporate private universities and non-tenured academics. Also, other aspects (number of patents, number of projects, number of spin-offs, etc.) will be used as variables in the cluster analysis.

Chapter 10

Effects of Research Collaboration

10.1 Introduction

Collaboration is a fundamental aspect of scientific research activity. It is considered the key issue for solving complex problems in many areas of science [109]. Generally, scientific collaboration could be defined as researchers working together to achieve the common goal of producing new scientific knowledge.

Collaboration usually helps researchers to share their workloads, generate fresh ideas, and combine peer past experience and skills [29, 199, 370]. These are all good reasons for collaboration, but they come at the expense of seeking the proper research partners, negotiating objectives, methodologies and results, managing geographic distance constraints, and communicating across organizations, cultures and disciplines and so on [38, 248, 274, 349].

Nowadays, researchers have begun to pay special attention to research performance and its determinants. Collaboration could be a determinant for achieving better research quality. Many researchers feel that collaborative research generally produces higher quality and more significant results than that performed by single researchers. They are motivated by the assumption that synergy leads to more and better results. A recent study [291] explains this point by arguing that each researcher has his own knowledge and the diversity of collaborating members could be an extra resource for reinforcing research quality.

Several bibliometric studies have explored the relationship of collaboration on the research performance. The relation between collaboration and productivity was first studied by Beaver and Rosen [39]. Authors concluded that collaboration is associated with higher productivity. Recently, Franceschet and Costantini [158] analyzed the relationship of scholar collaboration on the impact and quality of academic papers. They noted a general positive association between the cardinality of a paper's author set and the citation impact and peer quality of the contribution. Other studies have also corroborated that research collaboration has a positive influence on the number of documents [364] and the number of citations [425].

The practice of collaboration, and especially international collaboration, is becoming a widespread phenomenon. Some studies have shown a constant increase in terms of the number of papers with international collaborations [23], and an exponential increase in terms

of the number of international addresses [362]. This co-authorship trend is not surprising since it is an important aspect of an ideal work environment and it is also receiving interest and stimulus from policy-makers. Recent studies have analyzed the link between degree of internationalization of scientific activity and research performance at the level of individual researchers [7, 8]. They concluded that the top-performing national researchers also collaborate more abroad, but the reverse is not always true. Other studies demonstrated that the number of documents and the number of citations are positively correlated to the degree of international collaboration by a researcher [179, 451].

It is well-known that collaboration varies across disciplines and countries. On the one hand, Gazni *et al.* [177] performed a large-scale analysis to examine collaboration differences across multiple areas and from all countries. They found that the level of scientific collaboration varies dramatically by discipline. The life sciences display high levels of co-authorship, whereas the social sciences show low levels of co-authorship. Their analysis of the collaborations between countries revealed that six countries (United States, United Kingdom, Germany, France, Italy, and Canada) account for 82% of the world's international publications, but they are not the most collaborative countries, if measured by their proportion of collaborative output. On the other hand, Lancho Barrantes *et al.* [273] explored the provenance of the citations received by the different countries and the different types of collaborative papers. They found different percentages of papers in collaboration among countries. They also found that there is no significant correlation between scientific production and percentage of collaboration of a country. However, there is a significant negative correlation between production and the percentage traffic of citations to/from the collaborating countries. Regarding collaborative papers, they also found that there is a negative correlation between a country's production and its impact on domestic papers per paper. Finally, Franceschet and Costantini [158] analyzed the intensity of research collaboration in different areas. They observed that collaboration is negligible in arts and humanities. They also found that the scale and formality of social science collaborations are smaller than in science disciplines. Focusing on science disciplines, collaborative work is heavily exploited in chemistry, physics, biology and medicine. In contrast, it is moderate in mathematics, engineering and computer science. Despite this, the computer science field has been expanding since 1960 in terms of both number of published papers and number of authors. Also, computer science collaborations among different research institutes and across different countries have grown considerably recently [157]. According to Fortnow [151], it is time for computer science to grow up: it is now a mature field, and no major university can survive without a strong computer science department.

Franceschet [157] studied collaboration in computer science by means of a network science approach. Using publications from the DBLP Computer Science Bibliography, he examined properties like authors' scientific productivity and level of collaboration on papers, as well as large-scale network properties (average separation distance among scholars, distribution of the number of scholar collaborators, and dependence on star collaborators, among others). Franceschet concluded that the collaboration level in computer science papers is rather

moderate (two or three authors) compared with other scientific fields. Also, he observed that the computer science collaboration network is a widely connected small world. Hence scientific information flows along collaboration links very quickly and potentially reaches almost all scholars in the discipline. Finally, he noted that the distribution of collaboration among computer science scholars is highly skewed and concentrated, where a star collaborators are responsible for a relatively high share of collaborations. Despite this, the network connectivity does not crucially depend on them. Like Franceschet, this chapter deals with bibliometric properties such as author productivity and level of collaboration on papers. Unlike Franceschet, it is included the number of citations and citations per document and year. This analysis focuses on analyzing not network properties, but other aspects like types of collaboration, computer science subdisciplines and journal impact factor quartiles.

This chapter, which is based on the published paper [220], analyzes the relationship among research collaboration, number of documents and number of citations of the computer science research. Mainly, the number of documents and citations by number of authors is analyzed. These measures are also analyzed (according to the author set cardinality) under different circumstances, that is, when documents are written in different types of collaboration (international, national and institutional), when documents are published in different document types (journal article and conference paper), when documents are published in different computer science subdisciplines (artificial intelligence, cybernetics, hardware and architecture, information systems, interdisciplinary applications, software engineering and theory and methods), and, finally, when documents are published by journals with different impact factor quartiles (first-quartile journals, second-quartile journals, third-quartile journals and fourth-quartile journals). Note especially that there are no studies in the literature that investigate relationships among the above issues. Therefore, the following relationships are investigated:

- *Author cardinality vs Documents vs Citations*: The percentage evolution over time of documents published by number of authors and the average number of authors per document are analyzed. It is also analyzed the number of citations per document and year according to the documents author set cardinality.

- *Author cardinality vs Documents vs Citations vs Types of collaboration*: In this case, it is analyzed the trend of documents published as a result of international, national and institutional collaboration by number of authors. The average number of authors per document is also analyzed according to different types of collaboration. Finally, citation measures of documents published as a result of international, national and institutional collaborations are also explored by number of authors.

- *Author cardinality vs Documents vs Citations vs Document type*: In this case, the trend of documents published as journal articles and proceeding papers is analyzed by number of authors. The average number of authors per document is also analyzed according to different document types. Finally, citation measures of documents published in journals and conferences are also explored by number of authors.

- *Author cardinality vs Documents vs Citations vs Subdisciplines*: It is explored how documents published in different computer science subdisciplines change over time according to number of authors. The average number of authors per document according to different computer science subdisciplines is also analyzed. Finally, the number of citations per document and year in documents published in different computer science subdisciplines is studied by author cardinality.

- *Author cardinality vs Documents vs Citations vs Impact factor*: The percentage trend of documents published in different journal impact factor quartiles is studied by number of authors. The average number of authors per document according to different journal impact factor quartiles is also analyzed. Finally, citation measures of the above documents are analyzed against author cardinality.

Chapter outline

Section 10.2 firstly describes the indicators and statistical tests used to analyze the effects of research collaboration. It also reports the questions, hypotheses and results investigated in this chapter. Section 10.3 presents a discussion and conclusions on the results and indicates possible future research directions.

10.2 Questions, hypotheses and results

To investigate the above relationships, it has been analyzed the publications produced by active Spanish university professors between 2000 and 2009, working in the computer science field. Using these publications, some indicators regarding quality, collaboration, internationalization are calculated by number of authors.

The number of documents and citations are indispensable for analyzing research activity. Citations are measures of information use, reception and, in a way, of influence [107]. They can be considered as an indirect measure of publications quality in most cases, although there may be retracted papers that receive a lot of citations. By collaboration, two measures which are generally used in studies of research collaboration [286] are also computed. They are the collaborative rate, which is the percentage of documents with more than one author, and the collaborative level which is the average number of authors per document. Regarding the measures of internationalization [8], the international rate is used to analyze the percentage of papers that have been produced in collaboration with foreign institutions, that is, the percentage of publications co-authored with at least one co-author from an foreign institution. This measure is computed by analyzing the publications whose affiliations include addresses from more than one country. Finally, the impact factor is used as the status of a journal for a specific year. It is still recognized as the primary measure of journal quality and has a major influence on scientific behavior [473]. Furthermore, experience has shown that the best journals in each specialty are the publications in which it is hardest to get an article accepted, and these are the journals that have a high impact factor [175].

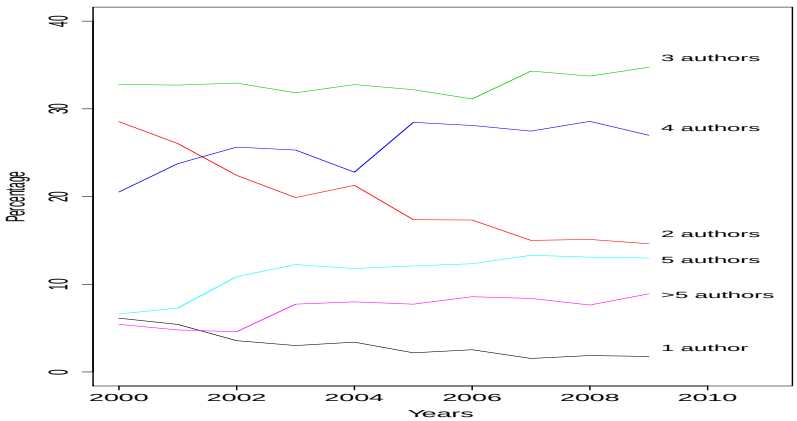
Once the above indicators are calculated, statistical tests are used to determine whether there is enough evidence to reject a conjecture about the data. The conjecture is called the null hypothesis. Not rejecting the conjecture may be a good result if it is wanted to continue to act as if the null hypothesis is true. Or it may be a disappointing result, possibly indicating that there is enough information to reject the null hypothesis. Tests that do not make assumptions about the population distribution are referred to as non-parametric tests. All commonly used non-parametric tests rank the outcome variable from low to high and then analyze the ranks. In this chapter, two non-parametric tests were used: Kruskal-Wallis test [269] and Mann-Whitney test [308]. The Kruskal-Wallis test analyzes whether three or more samples could have come from the same distribution. The null hypothesis is that the populations from which the samples originate have the same distribution. When the Kruskal-Wallis test leads to significant results, then at least one of the samples is different from the other samples. The test does not identify where the differences occur or how many differences actually occur. In contrast, the Mann-Whitney test analyzes whether two samples could have come from the same distribution. It is helpful for analyzing the specific sample pairs for significant differences. The significance level of these tests was 0.05 in all cases.

The following sections analyze the relationship among documents, citations and author cardinality on several issues such as types of collaboration, document types, computer science sub-disciplines and journal impact factor quartiles. The number of authors has been grouped into six different subsets (1 author, 2 authors, 3 authors, 4 authors, 5 authors and >5 authors).

10.2.1 How do productivity and visibility vary according to the number of authors?

The first question investigates the number of documents and citations regarding the author cardinality. The first impression is that computer science documents are usually the result of collaboration. Specially, it is expected that the average document is written by three or four authors. This is based on the idea that different co-authors reinforce research quality. It is

Figure 10.1: Evolution of percentage of published documents by number of authors



also expected that the number of authors per document has gradually increased in the last decade. Regarding visibility, it is supposed that a greater number of authors can lead to a higher number of citations because co-authors are more likely to disseminate the document.

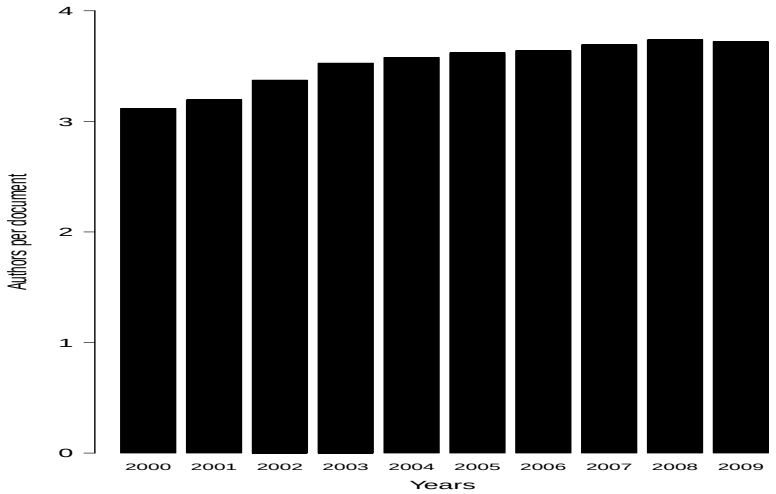
According to the different author subsets, document distribution in the analyzed period was: 1 author (2.651%), 2 authors (18.182%), 3 authors (33.037%), 4 authors (26.456%), 5 authors (11.994%), and >5 authors (7.680%). Most documents were published by three and four authors, whereas single-authored documents accounted for the lowest percentage.

Analyzing the number of authors, Figure 10.1 plots the evolution of documents published from 2000 to 2009. It shows that the percentage of documents published by different authors underwent some changes in the last decade. In earlier years, documents published by two authors accounted for a percentage of total publications (28.538% in 2000), but in 2009, it represented 14.616% of total publications. In contrast, the percentage of documents with three or more authors increased. The percentage of documents published by one author also decreased over the analyzed years, and, therefore, the collaborative rate increased over time.

As expected, these results bear out previous works stating that the practice of collaboration is becoming a widespread phenomenon. The number of authors used to be lower than it is today. Just a few authors were responsible for the hypothesis, experimental design, results and conclusion [488]. Nowadays, most projects require the participation of many researchers, who are all entitled to be authors when the results are reported. Other reasons that have increased the number of authors per document in recent years are dependency on the department chair and the addition of influential authors to raise a paper's prestige, among others. These authors who are neither author neither contributors are called guests [272, 424].

Figure 10.2 shows the evolution of the average number of authors per document. Collaborative level has increased in the last few years. Values rose from 3.118 authors in 2000 to 3.739 authors in 2008, so the increase was 19.917%. Taking the 2009 year as an example publication year, the published documents had an average of 3.721 authors per document.

Figure 10.2: Evolution of the average number of authors per document



Analyzing these values, the impression is that documents published by three or four authors will be the trend in computer science literature in the coming years.

Table 10.1 shows the average number of citations, the average number of years since the publication year, and the average number of citations per document and year (columns). These measures and their standard deviations are calculated for each different number of authors (rows). Analyzing Table 10.1, it is observed that the highest average number of citations (3.019±7.260) corresponded to documents published by one author, whereas the lowest average value (1.852±4.830) corresponded to documents published by five authors (column 2). These results were influenced by publication age, that is, the number of years since the publication year. The average age (column 3) is calculated to account for the above point. In this way, documents published by one author had the highest average age (5.536±2.821), whereas documents published by more than five authors had the lowest average age (4.172±2.460). The number of citations per document and year was calculated as an accurate measure for comparing documents published by number of authors. This ratio, which is a visibility measure is a possible indirect measure of the document's quality. In this context, documents published by two authors had the highest value (0.478±1.293), and documents published by five authors had the lowest value (0.363±0.841). Note that documents published by one or two authors had higher values of citations per document and year than documents published by three or more authors. A possible explanation could be that an important percentage of documents published by one or two authors are review papers. A review paper is usually written by a single senior research, and it is likely to be cited extensively. This would explain why single-authored documents received more citations than multi-authored documents.

The results of the Kruskal-Wallis test showed that there were significant differences among the six author subsets on the basis of average number of citations per document and year. So, Mann-Whitney tests were run in order to find out which subsets rank better according to this criterion. Documents published by two authors (benchmark subset), which had the highest average value, were compared with the other documents. Subsets marked in Table 10.1 with the symbol † had statistically significant differences with respect to the benchmark subset (highlighted in boldface). Results show that there were significant differences between the 2-author subset and subsets with more authors. Unlike Franceschet and Costantini [158], a positive association between the author set cardinality of a document and citation impact was not found in this analysis.

Table 10.1: Mean ± standard deviation of citation measures for documents published by different number of authors. The symbol † represents those results that are statistically different in citations ratio with respect to the benchmark subset (highlighted in boldface).

Authors	Citation count	Publication age	Citations ratio
1	3.019±7.260	5.536±2.821	0.443±0.889
2	2.943±10.297	5.023±2.733	**0.478±1.293**
3	2.181±6.562	4.450±2.634	0.407±1.021 †
4	1.882±5.469	4.346±2.548	0.365±0.903 †
5	1.852±4.830	4.200±2.453	0.363±0.841 †
>5	1.913±5.913	4.172±2.460	0.409±1.090 †

10.2.2 How do productivity and visibility in different types of collaboration vary according to author cardinality?

The second question analyzes whether productivity and visibility behave differently across different types of collaboration. A distinction between three types of collaboration is made: international, national and institutional cooperation. International collaboration refers to co-authorship by researchers from both national and foreign institutions. National collaboration refers to co-authorship by researchers belonging to different institutions in the same country. Finally, institutional collaboration refers to co-authorship among researchers belonging to the same institution.

Due to problems of geographic distance and communication across organizations, it is expected that most documents are written through institutional collaboration and there are more authors per document resulting from international collaboration than via national and institutional collaboration. On the other hand, analyzing visibility across different types of collaboration, it is reasonable to expect, precisely because of the differences among authors, the quality of documents resulting from international collaborations to be greater, and have a higher number of citations. Also, it is supposed that a greater number of authors can lead to a higher number of citations for a particular type of collaboration.

The document distribution in the studied period was: international collaboration (13.334%), national collaboration (13.112%) and institutional collaboration (73.554%). So, the value of the international rate was 13.334%. This percentage was very similar on a year-by-year basis. Therefore, the evolution of the international rate has not undergone major changes in the analyzed period. Results show that most collaborative documents were published via institutional collaboration.

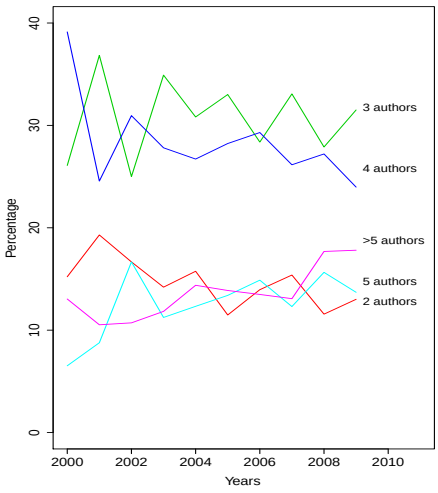
Figure 10.3 represents the evolution of published documents by number of authors and type of collaboration. Regarding international collaboration, Figure 10.3a shows that the percentage of documents published by number of authors has recently undergone changes. Result show that documents published by three and four authors represented the highest percentages of published documents each year, whereas documents published by two, five, and more than five authors represented the lowest percentages. Analyzing Figure 10.3a, it is observed a sizeable decrease in the percentages associated with documents published by two and four authors (e.g., the percentage of documents published by four authors was 39.130% in 2000 and 24.476% in 2009). In contrast, the number of documents published by three authors fluctuated considerably, and there were increases in the number of documents published by five or more authors. Regarding these increases, percentages associated with documents with five authors rose from 6.522% in 2000 to 16.776% in 2008, and percentages associated with documents with more than five authors rose from 10.526% in 2001 to 18.182% in 2009.

By national collaboration (see Figure 10.3b), results show an important decrease of documents published by two authors. The percentages associated with these documents were 19.022% in 2004 and 7.189% in 2009. Likewise, the percentage of documents with three authors also decreased from 41.818% in 2000 to 30.719% in 2009. In contrast, percentages associated with documents published by four or more authors increased over the time period.

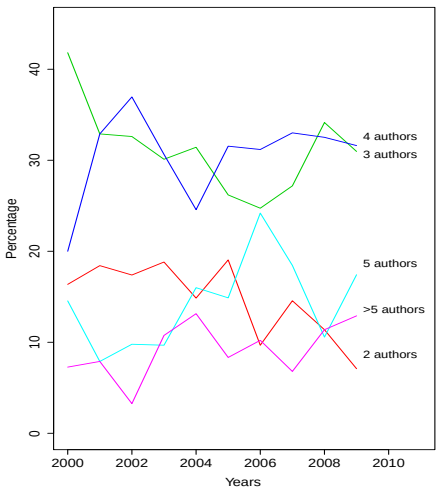
Figure 10.3c analyzes institutional collaboration. A sharp decrease in the percentage of documents published by two authors is observed. In earlier years, collaboration between two authors represented a sizeable percentage (35.353%) of the total publications, but this percentage decreased considerably (16.101%) in 2009. The other percentages associated with documents with three or more authors have gradually increased over last few years.

Finally, note that publication behavior has been similar across different types of collaboration in recent years. There were two different groups: documents published by three or four authors which had the highest percentages, and documents published by two, five, or more than five authors that had the lowest percentages. These groups were also highlighted

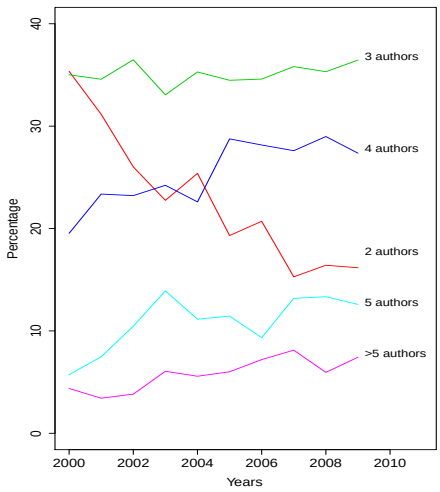
Figure 10.3: Evolution of percentage of published documents by number of authors and type of collaboration (international, national and institutional)



(a) International collaboration

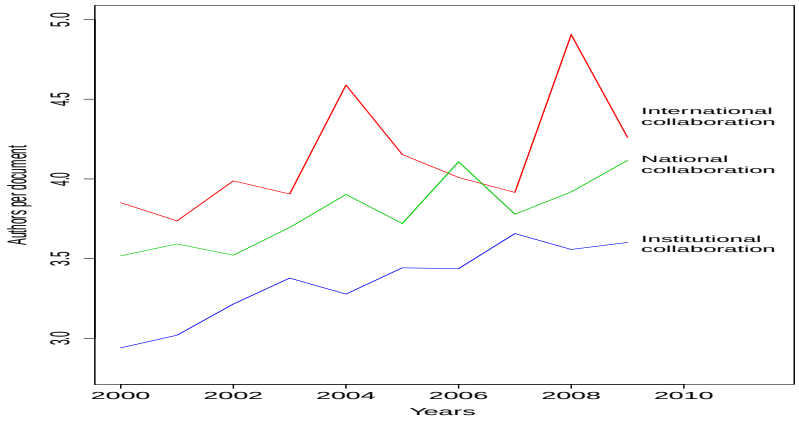


(b) National collaboration



(c) Institutional collaboration

Figure 10.4: Evolution of the average number of authors per document according to different types of collaboration (international, national and institutional)



in Figure 10.1 plotting the evolution of the percentage of published documents by number of authors. Figure 10.1 also shows an important decrease of documents published by two authors via international, national and institutional collaborations.

Regarding collaborative level, Figure 10.4 shows the average number of authors per document for each type of collaboration. Taking 2009 as an example publication year, it is observed that the average international document was published by 4.273 authors, whereas the average national document was published by 4.111 authors, and the average institutional document was published by 3.603 authors. According to the above values and the evolution illustrated in Figure 10.4, international collaborations usually had the highest number of authors per document, followed by national and institutional collaborations, as expected. A large number of international and national collaborations spring from projects that require the participation of many researchers from different institutions, whereas most institutional collaborations usually involve authors from the same research group. For these reasons, both international and national collaborations involve more authors than institutional collaborations, increasing the number of authors per document.

Table 10.2 shows the average number of citations, the average number of years since the publication year, and the average number of citations per document and year of documents published via different types of collaboration (international, national and institutional) and written by different numbers of authors. It also shows the standard deviations associated with the above measures. Note that international collaborations usually had the highest average values of citations per document and year for different numbers of authors, followed by national collaborations and institutional collaborations. International collaboration often involves more authors than other types of collaboration as mentioned before. As the authors are likely to disseminate the document, it is reasonable to assume that there will be a greater number of citations. Taking documents published by more than five authors as an example, note that the average values of international, national and institutional documents were 0.837 ± 1.816, 0.506 ± 0.804 and 0.207 ± 0.607, respectively. Like Glanzel [179] and

Table 10.2: Mean ± standard deviation for citation measures of international, national and institutional collaborations in documents published by number of authors. The symbol † represents those results that are statistically different in citations ratio with respect to the benchmark subset (highlighted in boldface)

	Collaborations		
	International	National	Institutional
2-authors			
Citation count	4.899±12.786	5.216±18.433	2.381±8.127
Publication age	4.974±2.487	5.584±2.384	4.955±2.800
Citations ratio	0.831±1.790	**0.749±2.066**	**0.395±1.046**
3-authors			
Citation count	3.765±6.813	3.995±8.608	1.620±6.022
Publication age	4.799±2.413	5.131±2.669	4.281±2.642
Citations ratio	0.716±1.182	0.722±1.218	0.305±0.932 †
4-authors			
Citation count	4.320±8.628	3.002±6.267	1.141±4.104
Publication age	4.909±2.453	4.931±2.534	4.104±2.532
Citations ratio	0.787±1.453	0.599±1.093	0.228±0.634 †
5-authors			
Citation count	3.552±6.858	2.682±5.201	1.285±4.028
Publication age	4.551±2.331	4.621±2.383	4.021±2.478
Citations ratio	0.671±1.146	0.573±1.047	0.245±0.662 †
>5-authors			
Citation count	3.750±9.821	2.579±4.078	0.980±3.588
Publication age	4.489±2.456	4.794±2.483	3.869±2.412
Citations ratio	**0.837±1.816**	0.506±0.804	0.207±0.607 †

Van Raan [451], these results demonstrate that, on an average, international collaboration results in documents with higher citation rates than national and institutional documents.

A Kruskal-Wallis test was performed in order to compare different subsets of authors (according to a particular type of collaboration). The Kruskal-Wallis test did not find statistically significant differences across international and national documents published by different authors. In contrast, results show that there were significant differences among institutional documents published by different authors. So, several Mann-Whitney tests were carried out to find out which subsets of authors (highlighted by †) were significantly different from the benchmark subset (highlight in boldface). It is demonstrated that institutional documents published by two authors were significantly different to all other subsets of authors. Analyzing the statistical test results, it is concluded that it is better to publish with few authors in order to improve document visibility at the institutional level, whereas the number of authors does not affect the average number of citations per document and year at the national and international level.

10.2.3 How do productivity and visibility in different document types vary according to author cardinality?

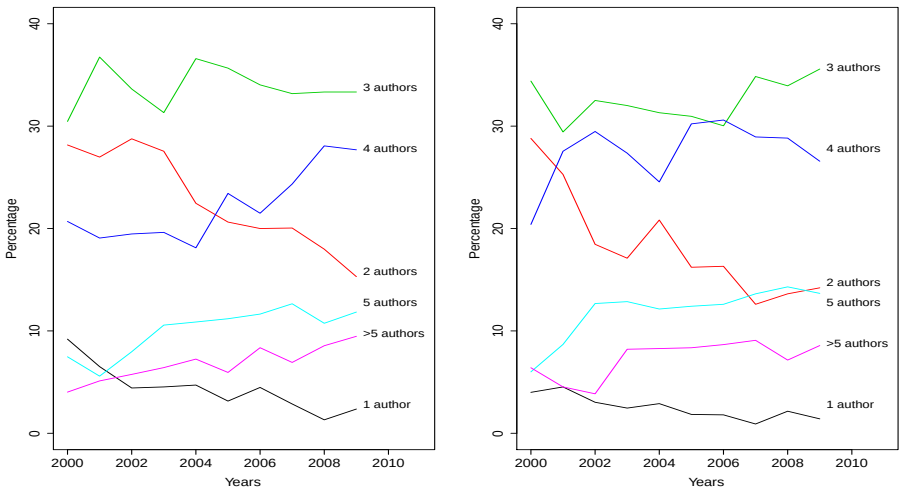
The third question analyzes whether productivity and visibility behave differently across different document types. Journal articles and conference papers are the document types studied in this section.

It is expected that publication behavior is different across journals and conferences. Due to the undeniable advantage of conferences (provide fast and regular publication of papers and bring researchers together by offering the opportunity to present and discuss the paper with peers), authors tend to publish more documents in conferences than in journals. It is also expected that most journal articles and conferences papers are published by three or four authors. By collaborative level, it is supposed that there are no clear differences between journals and conferences. On the other hand, it is expected that citation counts received by journal articles are higher than received by conference papers because of their prestige, and multi-authored documents receive more citations than single-authored documents.

The document distribution in the analyzed period was: journal articles (32.262%) and conference papers (67.738%). These percentages bear out previous works, like Franceschet [156], stating that 1/3 of computer science literature are journal articles and 2/3 are conference papers. These percentages of journal and conference documents usually vary on a year-by-year basis. Result also show that the percentage of conference papers have gradually decreased. In 2005, conference papers accounted for a sizeable percentage of total publications (74.000%), but in 2009, it represented 62.678% of total publications. An interpretation could be that researchers are progressively shifting from conferences to journals, considering budget shortages or higher prestige of journals over conferences.

According to the number of authors, 54.098% of single-authored documents are published in journals, whereas 45.902% are published in conferences. The rest of percentages were: 2 authors (43.098% in journals and 56.902% in conferences), 3 authors (36.906% in journals and 63.094% in conferences), 4 authors (32.839% in journals and 67.161% in conferences), 5 authors (31.750% in journals and 68.250% in conferences), >5 authors (34.002% in journals

Figure 10.5: Evolution of percentage of published documents by number of authors and document type



(a) Journal articles

(b) Conference papers

and 65.998% in conferences).

Figure 10.5 shows the evolution of the percentage of documents published in computer science journals and conferences by number of authors from 2000 to 2009. Result show that the percentage of documents associated with each author subset was similar in journal articles and conference papers, so there are no important differences in publication behavior by number of authors between journals and conferences. In general, there was a decrease in the number of documents published by one and two authors in both cases. Also, documents written by three and four authors accounted for the highest percentages, whereas the lowest percentage of documents were written by one author. Taking the journal articles as an example, Figure 10.5a shows that the percentage of documents with four or more authors has gradually increased over the last few years. Specially, documents published by four authors have undergone an increase in the last few years, they accounted for 18.116% of all publications in 2004, and 27.687% of total publications in 2009. In contrast, documents published by one author and two authors have decreased over the analyzed years and single-authored documents account for the lowest percentage in the 2002-2009 period.

Regarding collaborative level, Figure 10.6 shows the average number of authors per document for journal articles and conference papers. According to its evolution, conference papers have had the highest number of authors per document in earlier years. Despite this, journal articles and conference papers had similar number of authors per document in recent years. Taking 2009 as an example publication year, the average journal article was published by 3.738 authors, whereas the average conference paper was published by 3.711 authors.

Table 10.3 shows the average number of citations, the average number of years since the publication year, and the average number of citations per document and year. These measures and their standard deviations are calculated for each different number of authors and document type. It is noted that documents published by more than five authors had the highest average value of citations per document and year (0.971±1.734) when they were published by journals. In contrast, single-authored documents had the highest average value

Figure 10.6: Evolution of the average number of authors per document according to different document types

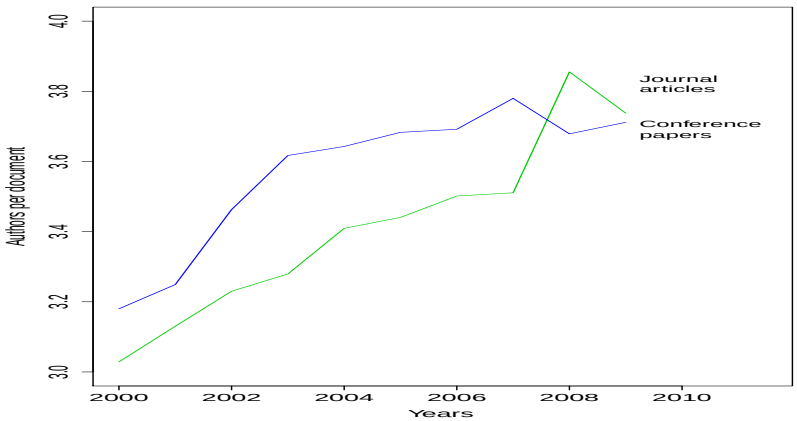


Table 10.3: Mean ± standard deviation for citation measures of journal and conference documents published by number of authors. The symbol † represents those results that are statistically different in citations ratio with respect to the benchmark subset (highlighted in boldface)

	Document type	
	Journal article	Conference paper
1-author		
Citation count	4.983±9.991	1.371±2.757
Publication age	5.783±2.937	5.329±2.713
Citations ratio	0.698±1.171	**0.229±0.456**
2-authors		
Citation count	6.029±15.774	1.063±3.133
Publication age	5.092±2.858	4.981±2.654
Citations ratio	0.940±1.917	0.196±0.495
3-authors		
Citation count	5.043±10.547	0.770±1.865
Publication age	4.534±2.794	4.408±2.551
Citations ratio	0.923±1.581	0.153±0.358 †
4-authors		
Citation count	4.753±9.045	0.746±2.205
Publication age	4.142±2.776	4.427±2.448
Citations ratio	0.924±1.457	0.143±0.354 †
5-authors		
Citation count	4.702±7.457	0.717±2.449
Publication age	4.118±2.603	4.233±2.392
Citations ratio	0.919±1.251	0.142±0.442 †
>5-authors		
Citation count	4.481±9.601	0.783±2.388
Publication age	3.974±2.657	4.259±2.366
Citations ratio	**0.971±1.734**	0.161±0.434 †

of citations per document and year (0.229±0.456) when they were published by conferences. As expected, journal articles had higher citations per document and year than conference papers. These results corroborate previous work, like Franceschet [156], in which the impact of journal publications was significantly higher than the impact of conference papers.

A Kruskal-Wallis test was performed in order to compare subsets of different authors across documents published in journals and conferences. Results show that there were no significant differences across documents published by journals. In contrast, it found significant differences across documents published by conferences: the average number of citations per document and year of documents published by one author (0.229±0.456) was significant different (higher) to documents published by three authors (0.153±0.358), four authors (0.143±0.354), five authors (0.142±0.442) and more than five authors (0.161±0.434).

10.2.4 How do productivity and visibility in different computer science subdisciplines vary according to author cardinality?

The fourth question investigates the productivity and visibility of authors across the seven *JCR* computer science subdisciplines: artificial intelligence, cybernetics, hardware and architecture, information systems, interdisciplinary applications, software engineering and theory and methods.

It is expected that publication behavior is different across subdisciplines. Authors should tend to publish more documents in mature disciplines like theory and methods. Also, the

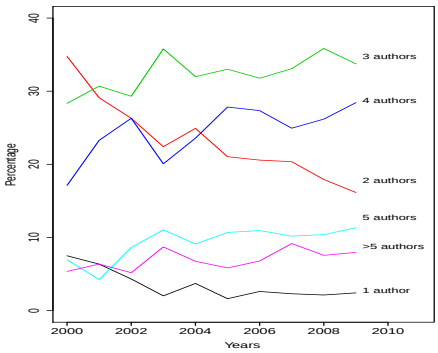
percentages of documents published by a specific number of authors should expect to be similar across subdisciplines. Furthermore, most documents should be published by three or four authors in all subdisciplines. Despite this, it is expected that the collaborative level is different. Interdisciplinary applications documents should be usually written by more authors than publications in other disciplines. This idea is based on the assumption that interdisciplinary applications documents could be published by authors belonging to many different areas, resulting in more authors per document. By visibility, it is expected that citation counts are different across subdisciplines and a greater number of authors leads to a higher number of citations in any particular a subdiscipline.

According to the Web of Science there is an overlap across the seven subdisciplines. Thus, one document could belong to more than one discipline at the same time. The document distribution in the analyzed period was: artificial intelligence (24.849%), cybernetics (1.613%), hardware and architecture (7.285%), information systems (9.528%), interdisciplinary applications (5.543%), software engineering (11.059%) and theory and methods (40.123%). Most documents were related to theory and methods, whereas cybernetics accounted for the lowest percentage of published documents.

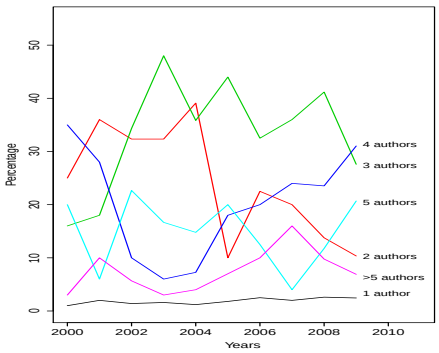
Figure 10.7 shows the evolution of the percentage of documents published in computer science subdisciplines by number of authors from 2000 to 2009. After analyzing all computer science subdisciplines, it is observed that the percentage of documents associated with each author subset was quite alike across different subdisciplines. These percentages were: 1 author [2.5274%-3.068%], 2 authors [18.001%-22.085%], 3 authors [32.594%-33.247%], 4 authors [24.773%-26.238%], 5 authors [9.811%-11.967%] and >5 authors [7.191%-8.333%]. According to these percentages, there were no important differences in publication behavior by number of authors across subdisciplines. Looking at all the charts illustrated in Figure 10.7, it is also observed similarities across subdisciplines: there was a general decrease in the number of documents published by one and two authors in all subdisciplines, documents written by three and four authors also accounted for the highest percentage in all subdisciplines, and the lowest percentage of documents were written by one author. On the other hand, the percentages associated with each subdiscipline have fluctuated widely in most computer science subdisciplines in the last decade. In contrast, artificial intelligence (see Figure 10.7a) and theory and methods (see Figure 10.7g) did not experience as many fluctuations as other subdisciplines. These two subdisciplines behaved very like computer science generally (see Figure 10.1). This was reasonable because these subdisciplines had the highest percentages of published documents, 24.849% and 40.123%, respectively.

Taking the artificial intelligence discipline as an example, Figure 10.7a shows that documents published by two authors accounted for the highest percentage (34.759%) of all publications in 2000, but represented only 16.145% of total publications in 2009. Documents published by one author have also decreased over the analyzed years and accounted for the lowest percentage in the 2002-2009 period. In contrast, the percentage of documents with three or more authors has gradually increased over the last few years.

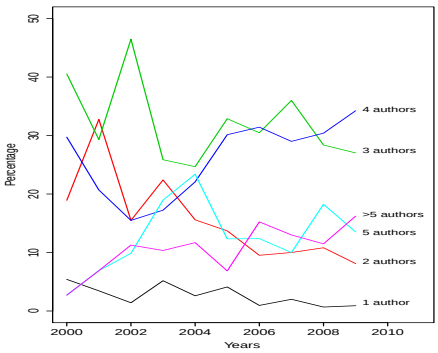
Figure 10.7: Evolution of percentage of documents published in computer science subdisciplines by number of authors



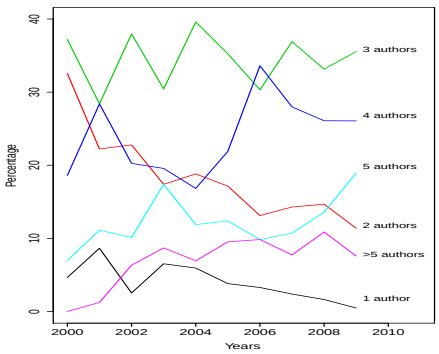
(a) AI: Artificial Intelligence



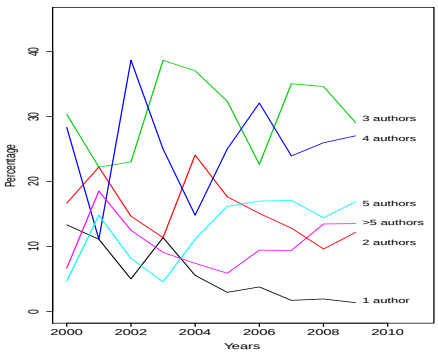
(b) CB: Cybernetics



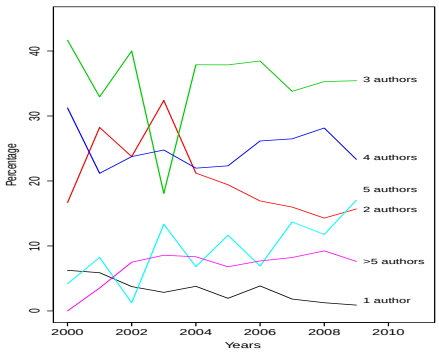
(c) HA: Hardware and Architecture



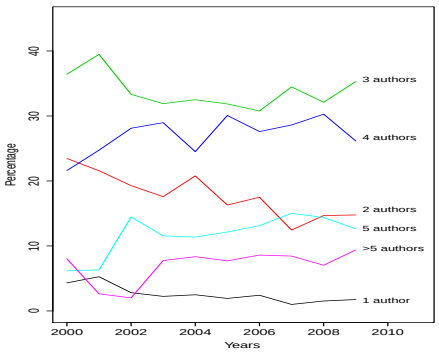
(d) IS: Information Systems



(e) IA: Interdisciplinary Applications



(f) SE: Software Engineering



(g) TM: Theory and Methods

Figure 10.8: Evolution of the average number of authors per document by disciplines

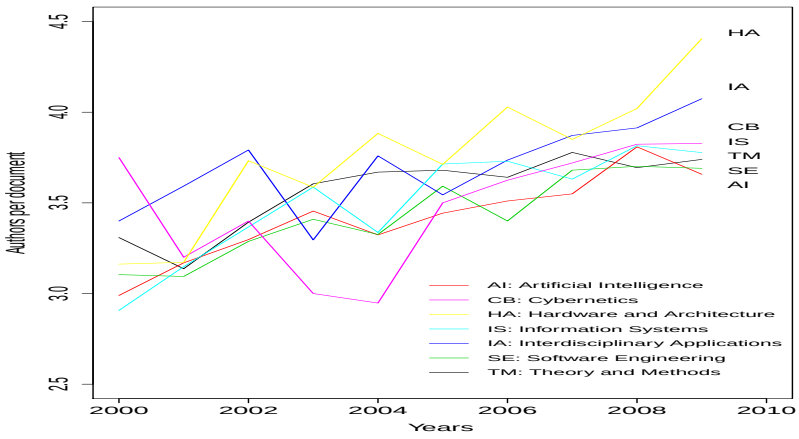


Figure 10.8 analyses collaborative level. It shows the evolution of the average number of authors per document according to different subdisciplines. These measures have tended to increase over the last few years. Note the hardware and architecture subdiscipline whose values rose from 3.162 in 2000 to 4.405 in 2009. In contrast, the number of authors per cybernetics document underwent a sizeable decrease up until 2004, and later recovered. Result also show that the range of the average number of authors per document was different across subdisciplines with respect to the analyzed year. Despite this, the range was wider in earlier years (2000-2004) than in later years (2005-2009). Finally, the highest values for collaborative level were achieved by documents belonging to the hardware and architecture (4.405 authors per document) and interdisciplinary applications (4.074 authors per document) subdisciplines in 2009. These values were the result of a major increment of documents published by more than three authors in these subdisciplines over the last few years (see Figure 10.7).

Table 10.4 shows the average number of citations, the average number of years since the publication year, and the average number of citations per document and year of documents published in different subdisciplines and written by different numbers of authors. It also shows standard deviations of the above measures. Analyzing the average number of citations per document and year, it is observed that some subdisciplines were more often cited than others. It is noteworthy that artificial intelligence documents, which had a lower value of authors per document than other subdisciplines, usually had a higher average values of citations per document and year than others. In contrast, hardware and architecture documents, which had the highest collaborative level value in recent years, received fewer citations than other subdisciplines like artificial intelligence, cybernetics and information systems. Citation counts by subdisciplines were known to vary within a particular discipline [50]. Some studies found that citation practices differ across subdisciplines. Like Smolinsky and Lercher [419], it is also found different citation behaviors by subdisciplines within a specific discipline (computer science in our case).

In order to compare citation behaviors by author subsets for a particular subdiscipline, several Kruskal-Wallis tests were performed. The Kruskal-Wallis test did not find meaningful differences across documents belonging to information systems, interdisciplinary applications and software engineering. In contrast, results show that there were significant differences among documents belonging to artificial intelligence, cybernetics, hardware and architecture, and theory and methods (see numbers in boldface and the † symbols). Taking artificial intelligence documents as an example, Table 10.4 shows that the number of citations per document and year of documents published by two authors (0.663±1.736) were significantly different to documents published by three (0.515±1.323) and four (0.551±1.283) authors. Similarly, the number of citations per document and year of hardware and architecture documents published by two authors (0.435±1.033) were significantly different to documents published by three (0.334±0.996) and four authors (0.229±0.641).

Analyzing the statistical test results in Table 10.4, it is concluded that the number of authors does not always affect the average number of citations per document and year. In this context, specific subdisciplines, like artificial intelligence, cybernetics, hardware and architecture, and theory and methods, are affected by the number of authors. Unlike Franceschet and Costantini [158], it is observed that documents with fewer authors usually have the highest average value of citations per document and year. Specifically, documents written by one author have the highest values for information systems, software engineering and theory and methods, whereas documents written by two authors have the highest values in artificial intelligence, cybernetics and hardware and architecture. In contrast, interdisciplinary applications documents published by more than five authors have the highest number of citations per document and year.

10.2.5 How do productivity and visibility in different journal impact factor quartiles vary according to author cardinality?

Journals ordered by impact factor can be organized into four quartiles. The first quartile denotes the top 25% of the impact factor distribution, the second quartile means a middle-high position (between the top 50% and top 25%), the third quartile is a middle-low position (top 75% to top 50%), and the fourth quartile represents a bottom position (bottom 25% of the impact factor distribution). The fifth question investigates the number of authors across different quartiles. Also, it is analyzed the productivity and visibility of documents published in different journal impact factor quartiles according to the author cardinality.

It is expected that first-quartile journals have the lowest publication rate due to their selective strategy, that is, low acceptance rates. Regarding the number of authors, the percentages of documents published by a specific number of authors should be similar across quartiles, so three or four authors per document should be the average collaborative level value. On the other hand, citation counts are obviously different across quartiles. Furthermore, it is expected multi-authored documents receive more citations than single-authored documents within a specific quartile.

Table 10.4: Mean ± standard deviation for citation measures of documents published by numbers of authors according to seven subdisciplines. The symbol † represents those results that are statistically different in citations ratio with respect to the benchmark subset (highlighted in boldface)

	AI	CB	HA	IS	IA	SE	TM
1-author							
Citation count	2.490±5.544	0.500±0.707	1.778±2.290	5.026±12.946	2.348±3.725	4.400±12.740	2.514±4.219
Publication age	5.792±3.067	2.500±2.121	5.944±2.689	5.949±2.502	5.632±2.790	5.714±2.771	5.180±2.728
Citations ratio	0.360±0.727	0.125±0.177	0.295±0.403	**0.625±1.483**	0.403±0.581	**0.557±1.383**	**0.427±0.653**
2-authors							
Citation count	4.263±13.641	7.952±21.971	2.838±8.624	2.335±7.479	1.826±4.359	1.649±4.062	2.451±9.990
Publication age	5.292±2.825	5.357±2.458	5.581±2.847	5.031±2.899	4.174±2.510	4.826±2.742	4.735±2.561
Citations ratio	**0.663±1.736**	**1.059±2.486**	**0.435±1.033**	0.479±1.272	0.363±0.747	0.291±0.639	0.394±1.184
3-authors							
Citation count	2.798±8.209	4.242±9.980	1.878±5.959	1.817±4.674	1.510±3.372	1.844±4.841	1.793±5.940
Publication age	4.622±2.702	4.323±2.289	4.776±2.793	4.301±2.722	3.784±2.451	4.221±2.694	4.397±2.549
Citations ratio	0.515±1.323 †	0.763±1.359	0.344±0.996 †	0.386±1.038	0.318±0.670	0.375±0.908	0.334±0.881 †
4-authors							
Citation count	2.743±7.271	2.657±5.886	1.307±4.694	2.108±6.930	1.476±4.821	1.754±4.484	1.337±3.871
Publication age	4.494±2.660	3.714±2.729	4.101±2.655	4.101±2.670	3.693±2.515	4.196±2.683	4.334±2.426
Citations ratio	0.551±1.283 †	0.731±1.802	0.229±0.641 †	0.399±0.995	0.338±0.777	0.338±0.783	0.258±0.639 †
5-authors							
Citation count	3.046±6.821	3.885±9.868	2.130±5.459	1.346±3.906	2.489±5.681	1.293±3.609	1.377±4.107
Publication age	4.375±2.571	4.500±2.746	4.200±2.436	3.949±2.716	3.298±2.233	3.487±2.432	4.120±2.327
Citations ratio	0.593±1.167	0.548±1.087	0.384±0.820	0.220±0.593	0.579±1.166	0.292±0.729	0.272±0.694 †
>5-authors							
Citation count	2.364±5.534	2.143±5.405	2.206±6.390	1.478±3.510	2.451±5.995	1.388±3.612	1.626±6.819
Publication age	4.467±2.669	3.643±2.307	3.979±2.504	3.696±2.208	3.549±2.666	3.825±2.341	4.069±2.354
Citations ratio	0.509±1.320	0.277±0.662 †	0.434±1.055	0.396±0.939	**0.776±2.090**	0.318±0.736	0.321±0.998 †

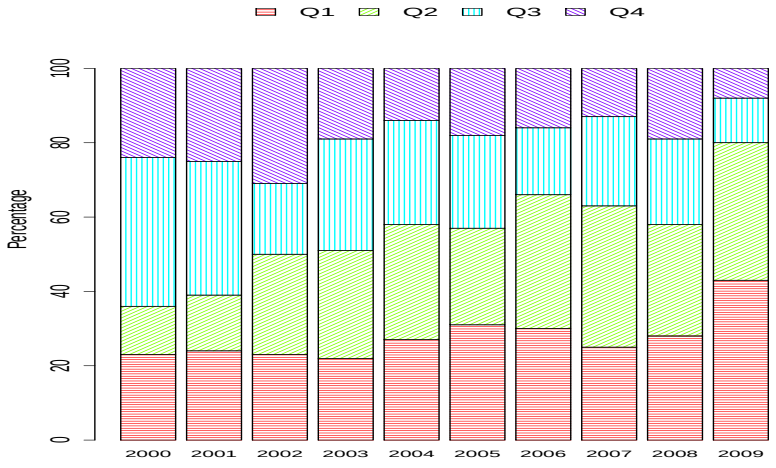
Table 10.5 shows the percentages of documents published in each quartile for different numbers of authors. In single-authored documents the percentages associated with each quartile were: first-quartile (28.333%), second-quartile (24.167%), third-quartile (31.667%) and fourth-quartile (15.833%). In this case, the third-quartile had the highest percentage of published documents. This quartile also had the highest percentage of documents published by two authors. In contrast, documents published by three or more authors were usually published in journals belonging to the first quartile. On the other hand, journals belonging to the fourth quartile accounted for the lowest percentages of published documents. Nowadays, authors have an interest in publishing in journals with the highest possible impact factor, and, therefore, it is reasonable to suppose that first-quartile journals accept more documents than fourth-quartile journals. This supposition bear out previous work, like Cabanac [72], stating that the range of papers accepted per journal is wider for the first-quartile than for the other quartiles.

According to the distribution of document published in different quartiles during the analyzed period, it is observed that first-quartile had the highest percentage (30.490%), followed by second-quartile (27.460%), third-quartile (27.335%) and fourth-quartile (14.715%). The evolution of documents published in different quartiles is analyzed in Figure 10.9. Note that the highest percentage of first-quartile documents were achieved in 2009, whereas the highest percentage of fourth-quartile documents were achieved in 2002. It is also observed that

Table 10.5: Percentages of documents published in each quartile by different numbers of authors

Authors	First-quartile	Second-quartile	Third-quartile	Fourth-quartile
1	28.333%	24.167%	31.667%	15.833%
2	27.086%	24.305%	33.821%	14.788%
3	30.037%	28.189%	27.726%	14.048%
4	32.392%	29.032%	23.387%	15.189%
5	31.268%	28.024%	23.304%	17.404%
>5	36.481%	29.185%	22.747%	11.587%

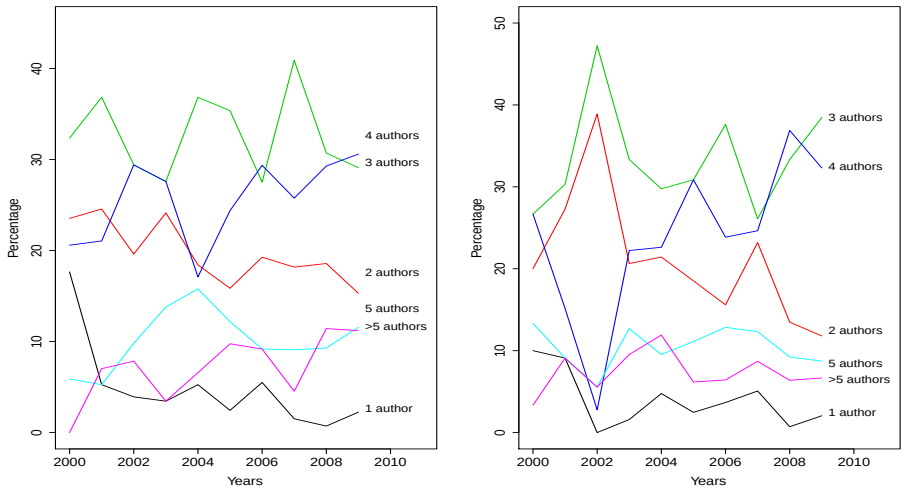
Figure 10.9: Evolution of percentages of documents published in quartile journals



the percentage of documents published by first- and second-quartile journals have gradually increased, whereas the percentage of documents published by third- and fourth-quartile journals have decreased.

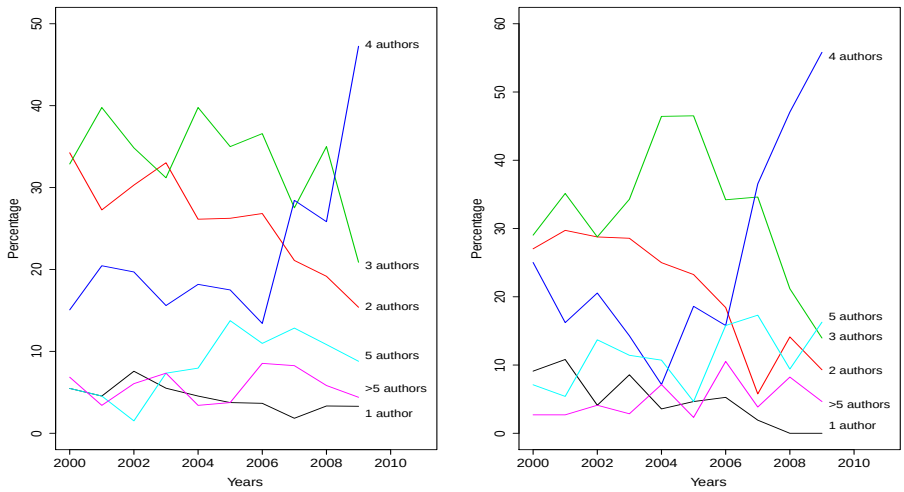
The evolution of documents published in different quartiles according to different numbers of authors is analyzed in Figure 10.10. It is found that there were small differences in publication behavior by author set cardinality for documents across different quartiles. In general, documents published by one and two authors have undergone a percentage decrease, leading to a drop in the collaborative rate value throughout the analyzed period. Also, the percentage of documents published by four authors has increased, whereas documents

Figure 10.10: Evolution of percentages of documents published in quartile journals by number of authors



(a) First-quartile

(b) Second-quartile



(c) Third-quartile

(d) Fourth-quartile

written by three, five and more than five authors have undergone fluctuations and small decreases with respect to different quartiles. As expected, documents published by three or four authors had the highest values in each quartile. There has been a noteworthy increment of documents published by four authors in last few years, rising to 47.252% of documents in third-quartile journals (see Figure 10.10c) and 55.814% of documents in fourth-quartile journals (see Figure 10.10d) in 2009. On the other hand, the main differences were associated with documents written by three authors. In this case, the percentages of these documents fluctuated in the first- and second-quartile journals, whereas they clearly decreased in third- and fourth-quartile journals.

Figure 10.11 analyzes the collaborative level with respect to different quartiles. It shows that fourth-quartile journals had the highest value for number of authors per document (4.519 authors) in 2009, followed by journals belonging to the first (3.869 authors), third (3.587 authors) and second (3.569 authors) quartiles. Also, all collaborative level values have undergone an increase in the last few years. Especially noteworthy was the sizeable increment of fourth-quartile journals since 2006. According to these results, fourth-quartile journals have recently accepted documents with many authors (see Figure 10.10d) in order to improve their number of citations. A possible interpretation is that these journals publish documents with many co-authors in order to improve their quartile through increased dissemination by co-authors including self-citations.

Table 10.6 presents the average number of citations per document, the average age per document, and the average number of citations per document and year. These values and their standard deviations were calculated for documents belonging to each quartile. As expected, documents published by first-quartile and second-quartile journals usually had a higher average number of citations per document and year than documents published by third-quartile and fourth-quartile journals. Analyzing the number of authors, it is noted that single-authored documents always had the lowest average value of citations per document and year when they were published by first-, second- and third-quartile journals. In contrast, these

Figure 10.11: Evolution of the average number of authors per document (CL) by journal impact factor quartiles

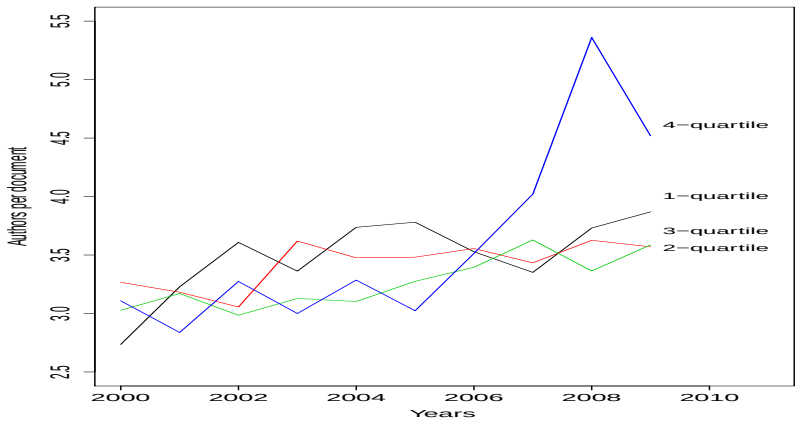


Table 10.6: Mean ± standard deviation for citation measures of documents published by numbers of authors according to impact factor. The symbol † represents those results that are statistically different in citations ratio with respect to the benchmark subset (highlighted in boldface)

	First-quartile	Second-quartile	Third-quartile	Fourth-quartile
1-author				
Citation count	6.176±11.862	2.069±3.240	4.237±7.713	8.789±15.183
Publication age	5.559±3.193	4.862±2.900	5.947±2.837	7.263±2.207
Citations ratio	0.815±1.255	0.460±0.816 †	0.574±0.938	**1.100±1.734**
2-authors				
Citation count	8.811±21.678	5.000±12.819	5.723±14.828	3.442±5.775
Publication age	4.227±2.837	4.530±2.625	5.792±2.763	6.253±2.737
Citations ratio	**1.474±2.717**	0.836±1.634 †	0.794±1.578	0.454±0.715
3-authors				
Citation count	6.708±13.177	3.859±7.928	5.683±11.643	2.736±3.679
Publication age	3.969±2.707	3.833±2.574	5.467±2.776	5.521±2.660
Citations ratio	1.359±2.167	0.767±1.219 †	0.860±1.380	0.452±0.589
4-authors				
Citation count	6.178±10.491	4.495±8.462	4.845±9.487	2.163±4.451
Publication age	3.859±2.701	3.662±2.413	4.891±2.945	4.750±3.052
Citations ratio	1.234±1.654	1.009±1.641	0.744±1.152	0.374±0.618
5-authors				
Citation count	4.991±7.321	5.684±8.714	5.000±7.402	2.389±4.866
Publication age	3.755±2.570	4.021±2.497	4.418±2.520	4.778±2.879
Citations ratio	1.037±1.265	**1.118±1.466**	0.936±1.209	0.390±0.632
>5-authors				
Citation count	5.176±8.181	3.853±4.643	5.962±16.233	1.042±2.053
Publication age	3.200±2.429	4.044±2.464	5.113±2.819	4.375±2.732
Citations ratio	1.153±1.465	1.029±1.959	**0.975±2.135**	0.238±0.488 †

documents had the highest average value (1.100±1.734 citations) when they were published in fourth-quartile journals. A Kruskal-Wallis test was performed in order to compare subsets of different authors across documents published in different quartiles. Results show that there were no significant differences across documents published by first- and third-quartile journals. In contrast, it found significant differences across documents published by second- and fourth-quartile journals: the average number of citations per document and year of documents published by five authors in second-quartile journals (1.118±1.466) was significant different (higher) to documents published by one author (0.460±0.816), two authors (0.836±1.634) and three authors (0.767±1.219). Likewise, the average number of citations per document and year of documents published by one author in fourth-quartile journals (1.100±1.734) was significant different (higher) to documents published by more than five authors (0.238±0.488). According to these results, no pattern is found to explain the relationship between impact factors, visibility and authors.

10.3 Discussion and conclusions

This analysis is limited to the Spanish computer science production included in the Web of Science. It is a small percentage of the worldwide output, therefore, the results may not be generally applicable. Further research is required in order to assess the above questions.

This chapter has studied five relationships. The first analyzes how productivity and visibility vary according by number of authors. According to productivity, the initial hypothesis

was that the average computer science document was written by three or four authors. The research findings confirm that the hypothesis was correct. Results also show that the collaborative level has increased over time as expected. This was caused by both the percentage decrease of published documents written by one and two authors, and the percentage increase of documents written by three or more authors. On the other hand, it was expected that a higher number of authors would lead to a higher number of citations. In contrast, results show that documents published by one or two authors have higher values of citations per document and year than documents published by three or more authors. In fact, statistical test results show that there are significant differences between the 2-author subset and subsets with more authors. A positive association between author set cardinality and the citation impact was not found.

The second relationship analyzes how productivity and visibility vary across different types of collaboration according to different number of authors. Due to the problems concerning geographic distance and communication across organizations, it was expected that most documents were written via institutional collaboration. Results show that the initial hypothesis was correct, since the 73.554% of total publications were published by authors belonging to the same institution. On the other hand, publication behavior was similar across different types of collaboration. International, national and institutional collaborations are usually written by three or four authors. Regarding collaborative level, the hypothesis was that international documents would have more authors than national and institutional documents. Results show that the initial hypothesis was again correct. A possible explanation of this fact is that a large number of international and national collaborations spring from projects that require the participation of many researchers at different institutions, whereas most institutional collaborations usually involve authors from the same research group, that is, involve fewer authors. Finally, it was also expected that international collaborations would have more citations than national and institutional documents. Unlike Bartneck and Hu [34] who were unable to find a general beneficial effect of collaboration of any type (international, national or institutional) on the quality of the papers measured by their citation counts, results show that international collaborations always have the highest average number of citations per document and year for different numbers of authors. It was also expected that a greater number of authors would lead to a higher number of citations with a particular type of collaboration. However, statistical test results show that document visibility is better if it is published by few authors at the institutional level, whereas the number of authors does not affect citation counts at national and international level.

The third relationship investigates how productivity and visibility vary across different document types according to different number of authors. It was expected that the publication rate associated with journal articles and conference papers would be different. Results corroborated this hypothesis, showing that 32.262% of publications belong to journal articles, whereas 67.738% of publications belong to conference papers. Analyzing the collaborative level, the initial hypothesis was correct. There are no important differences in publication behavior between journals and conferences as believed, but the collaboration level in confer-

ence papers was higher than journal articles in earlier years. Results also show that there is a general decrease of documents published by one and two authors in journal and conference publications. According to the number of authors, single-authored documents are usually published in journals, whereas multi-authored documents are usually published in conferences. It was also expected that citation counts would be different between journals and conferences. Results show that journal articles have more citations per document and year than conference papers. Finally, unlike the initial hypothesis, statistical test results do not assure that a greater number of authors leads to more citations.

The fourth relationship investigates how productivity and visibility among the seven computer science subdisciplines vary by number of authors. The initial hypothesis was that the publication rate associated with each subdiscipline would be different. Results corroborated this hypothesis, showing that 40.123% of publications belong to theory and methods, whereas only 1.613% of publications belong to cybernetics. Regarding the percentages of documents published by different numbers of authors, it is observed that there are no important differences in publication behavior across subdisciplines as believed. Results show that there is a general decrease of documents published by one and two authors in all subdisciplines. Also, three and four authors write the highest percentage of documents in all subdisciplines, whereas documents written by one author account for the the lowest percentage. Analyzing the collaborative level, the initial hypothesis was correct. It was also expected that citation counts would be different across subdisciplines. Results show that documents related to artificial intelligence, cybernetics and interdisciplinary applications usually have the highest value of citations per document and year with a set number of authors. Finally, unlike the initial hypothesis, statistical test results do not assure that a greater number of authors leads to higher number of citations within a specific a subdiscipline.

The last relationship analyzes how productivity and visibility in different journal impact factor quartiles vary by number of authors. It was expected that first-quartile journals would have the lowest percentage of publications due to their low acceptance rate. Contrariwise, results show that first- and second-quartile journals publish more documents than third- and fourth-quartile journals. This is reasonable bearing in mind that authors have an interest in publishing in journals with the highest possible impact factor nowadays. Regarding the number of authors, it was supposed that there would be no clear differences across quartiles. In contrast, results show an important increment of number of authors per documents in fourth-quartile journals since 2006. As expected, results show that citation counts are obviously different across quartiles. Finally, it was also expected that a greater number of authors would lead to a higher number of citations for a set quartile. However, statistical test results found no pattern to explain the relationship between impact factor quartiles, number of citations and number of authors.

In the future, the target will be to analyze other aspects related to collaboration at author level (number of different co-authors, productivity of co-authors, visibility of co-authors, proximity among co-authors, etc). Furthermore, it is interesting to analyze whether researchers with the best research performance are also the investigators that collaborate more at the

international level, and whether the citation counts of papers that have been written by authors with a low number of citations improve through collaboration.

Chapter 11

Predicting Spanish h-index

11.1 Introduction

One of the most successful bibliometric indices is the *h-index* [207], which quantifies the scientific output of a single researcher as a single-number criterion taking into account both the quantity and visibility of publications. Formally, the *h-index* is defined as follows: "A scientist has index h, if h of his or her N_p papers have at least h citations each and the other (N_p - h) papers have *leq* h citations each". It has acquired such significance in the evaluation of research over recent years. In fact, many funding agencies and promotion committees use it for accepting research projects and contracting researchers, among others.

Nowadays, many researchers have turned their attention to the prediction of bibliometric indices due to increasing interest within the scientific community. For example, Krampen *et al.* [268] forecasted the *number of documents* that a researcher would produce within ten years in the field of psychology. They used time series modeled by exponential and exponential smoothing functions. The predictions were based on past psychology publication frequencies. Some other works have predicted the *h-index* values under different circumstances. The power law model [136] was used to analyze the *h-index* as a function of time [133]. Nonlinear regression was also used to predict the *h-index* of authors, journals and universities [482]. Most of works concerned with predicting the *h-index*, only used *h-index* sequences to indicate by extrapolation what the value of the *h-index* would be in the near future. Recently, Acuna *et al.* [9] also used current *h-index* values to predict the *h-index* using a linear regression with elastic net regularization. Their model also included other predictors like number of articles written, number of years since publication of the first article, number of distinct journals in which the articles came out, and number of articles published in five top journals. Finally, McCarty *et al.* [312] predicted the author *h-index* using characteristics of the co-author network. Results of their regression model suggest that the highest *h-index* will be achieved by working with many co-authors, at least some with high *h-indexes* themselves.

The interest and originality of this chapter is a new approach based on cost-sensitive naive Bayes models for predicting the *h-index* for a four-year time horizon using some author-based variables (*area*, *position*, *university*, *seniority*) and 12 bibliometric indices. Specifically,

several prediction models are built to forecast the annual increase of the *h-index* of Spanish professors. Finally, the *h-index* prediction is considered as an ordinal classification problem. For this reason, it is not interested in maximizing the classification accuracy, but in minimizing the expected total cost error derived from mistakes in the classification process. This work appears in the published paper [224].

Chapter outline

Section 11.2 explains the proposed classification method (Section 11.2.1), the selection of important predictive features (Section 11.2.2) and the different models learned (Section 11.2.3). Finally, Section 11.3 contains some conclusions emphasizing the original contribution of this work and future research on the topic.

11.2 Predicting the h-index's annual increase

In view of the attention attracted by the *h-index*, this chapter is focused on the construction of several prediction models to forecast the *h-index* annual increase of Spanish professors for a four-year time horizon. Two different types of models (senior models and junior models) are learned to differentiate between professors' seniority. On the one hand, senior models attempt to predict the annual increase of the *h-index* of professors who had a seniority of at least eight years at the end of the information collection process (December 31, 2005), that is, they published their first paper before 1998. On the other hand, junior models also attempt to predict the annual increase of the *h-index*, but, in this case, only professors who had a seniority of at most three years at the end of the information collection process, were considered. These models are learnt from bibliometric data using a cost-sensitive naive Bayes approach that takes into account the expected cost of instances predictions at classification time.

11.2.1 Cost-sensitive naive Bayes approach

The cost-sensitive naive Bayes is an adaptation of the original naive Bayes [325] which is one of the simplest models for supervised classification. It is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. A cost-sensitive naive Bayes classifier has two types of variables: the class variable C and a set of predictive features $\mathbf{X}=\{X_1, X_2, ..., X_n\}$. The class variable C is discrete and takes values in the set $\Omega(C)$. The predictive features can be divided into two sets: the set of discrete features $\{X_1, ..., X_m\}$ and the set of continuous features $\{X_{m+1}, ..., X_n\}$. This classifier is based on Bayes' theorem under the assumption of conditional independence of predictor features given the class variable. The objective of the cost-sensitive naive Bayes is to take into account misclassification costs different from 0 (hit) and 1 (miss). Given a cost matrix and a set of predicted class probabilities for each instance, this method readjusts the probability thresholds of each class to select the class with the minimum-expected cost. The expected

cost of each prediction is obtained by multiplying the associated costs by the predicted class probabilities. The cost matrix is ignored when making predictions, but taken into account for their evaluation. Unlike the original naive Bayes, this method does not select the most likely class value of the posterior distribution, it selects the class (c^*) that minimizes the expected cost of predictions given a new instance $\mathbf{x}$:

$$c^* = \arg \min_{c \in \Omega(C)} \sum_{c'=1}^{val(C)} p(c' \mid \mathbf{x}) \;\; cost(c \mid c')$$

where

$$p(c' \mid \mathbf{x}) \propto p(c') \prod_{i=1}^{m} p(x_i \mid c') \prod_{j=m+1}^{n} \mathcal{N}(x_j, \mu_j^c, (\sigma_j^c)^2)$$

and $cost(c \mid c')$ is the associated misclassification cost.

11.2.2 Selecting predictive features

The objective of feature selection is to build parsimonious models. Features that are irrelevant or redundant will not appear in these models. The benefits of applying feature selection include better classification performance, faster classification models, smaller databases, and the ability to gain more insight into the process that is being modeled [390].

The predictive features used by cost-sensitive naive Bayes were:

- The feature *Area* represents the subject area related to each professor. It has three possible values: Computer Architecture and Technology, Computer Science and Artificial Intelligence, and finally, Computer Languages and Systems.

- The feature *Position* corresponds to the position of each professor, having four possible values: Full Professor, Associate Professor (type I), Associate Professor (type II) and Associate Professor (type III).

- The feature *University* is associated with the public university employing each professor. It has 48 possible values (e.g., Technical University of Madrid, University of Granada, University of Almeria, University of Castilla-La Mancha,...).

- The feature *Seniority* represents the seniority associated with each professor. It is a numeric feature which is calculated in terms of the publication year of his or her first paper.

- The bibliometric indices selected as predictive features were: *documents*, *citations*, the *h-index*, the *g-index*, the *hg-index*, the *a-index*, the *m-index*, the q^2*-index*, the h_r*-index*, the h_i*-index*, the h_c*-index*, and the *c-index*.

In this context, correlation-based feature selection is used as the feature selection algorithm. The basic idea behind this algorithm is to find a good set of features that are highly correlated with the class to be predicted and uncorrelated with each other.

11.2.3 Junior and senior predictive models

Once the dataset construction is finished, Table 11.1 shows the distribution of the professors selected in junior and senior models according to their annual increase of the *h-index* value within the first four years. Taking the first year as an example, it is observed that most junior professors have Δ*h-index*=0, whereas only few professors have Δ*h-index*=1. It is also showed that most junior professors have Δ*h-index*=0 in the second, third and fourth year. In contrast, senior models show different data distributions. For example, the *h-index* value for senior professors increases from 0 to 4 in the first year and from 0 to 14 in the fourth year.

Before learning the predictive models, a feature selection is performed in order to determine if all the predictive features are equally important or necessary for discriminating between the different values of the annual increase of the *h-index*. Table 11.2 shows the predictive features that are selected in senior models after running correlation-based feature

Table 11.1: Data distribution of junior and senior professors

Δh	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Junior professors															
First-year	239	50	-	-	-	-	-	-	-	-	-	-	-	-	-
Second-year	205	82	2	-	-	-	-	-	-	-	-	-	-	-	-
Third-year	159	119	10	1	-	-	-	-	-	-	-	-	-	-	-
Fourth-year	146	125	17	1	-	-	-	-	-	-	-	-	-	-	-
Senior professors															
First-year	267	76	7	1	1	-	-	-	-	-	-	-	-	-	-
Second-year	203	122	29	5	1	1	1	-	-	-	-	-	-	-	-
Third-year	165	118	54	9	4	0	0	1	0	0	0	1	-	-	-
Fourth-year	147	113	59	22	4	5	0	0	0	1	0	0	0	0	1

Table 11.2: Selecting predictive features of senior models

Features	First-year	Second-year	Third-year	Fourth-year
area				
position				
university	√	√	√	√
seniority				
documents	√	√	√	√
citations	√			√
h-index		√		
g-index	√	√	√	√
hg-index	√			
a-index				
m-index	√			
q^2*-index*		√	√	
h_r*-index*	√			
h_i*-index*		√		
c-index	√	√	√	√
h_c*-index*	√	√		

selection. This table illustrates that *university*, *documents*, *g-index* and *c-index* are always chosen. These predictive features are highly correlated with the class to be predicted. On the other hand, *area*, *position*, *seniority* and *a-index* are never selected to build parsimonious models.

Different cost-sensitive naive Bayes models are learned with the intention of checking the benefits of applying feature selection (e.g. better classification performance). The first model approach is learnt from a dataset (Dataset_{nofs}) which does not include the above feature selection whereas the second model approach is learnt from dataset (Dataset_{fs}) which does. The required cost matrix is associated with an exponential function. In this way, instances, whose weighted distance between the actual and the predicted class values is very high, will be heavily penalized. Also, k-fold cross-validation is chosen as the procedure for estimating the accuracy of models classifying new cases according to the value of the predictive features.

Table 11.3 lists the accuracy and the standard deviations of predictive models. It also shows the number of values of the class variable and the number of predictive features (between parentheses) accounted for by each model. Taking the prediction values of senior models for the first year as an example, it is noted that the class variable has 5 possible Δh values (0,1,2,3,>3) which are forecast using sixteen predictive features (Dataset_{nofs}) and nine

Table 11.3: Accuracy, standard deviations and number of features of models which are learnt from two different datasets

	Junior Models	Senior Models
First-year	2 classes	5 classes
$Dataset_{nofs}$	74.75 ± 12.05 (16 features)	68.85 ± 5.78 (16 features)
$Dataset_{fs}$	81.31 ± 2.37 (1 feature)	69.52 ± 5.58 (9 features)
Second-year	3 classes	7 classes
$Dataset_{nofs}$	47.74 ± 13.96 (16 features)	57.15 ± 6.68 (16 features)
$Dataset_{fs}$	71.46 ± 5.72 † (2 features)	58.20 ± 6.78 (8 features)
Third-year	4 classes	12 classes
$Dataset_{nofs}$	54.08 ± 8.23 (16 features)	53.58 ± 7.81 (16 features)
$Dataset_{fs}$	55.02 ± 6.94 (2 features)	51.02 ± 6.49 (5 features)
Fourth-year	4 classes	15 classes
$Dataset_{nofs}$	48.92 ± 7.23 (16 features)	47.85 ± 8.19 (16 features)
$Dataset_{fs}$	50.21 ± 7.90 (3 features)	48.51 ± 7.35 (5 features)

Table 11.4: Accuracy, standard deviations and average cost of models which are learnt using different naive Bayes approaches

	Junior Models	Senior Models
First-year	2 classes	5 classes
NB	81.31 ± 2.37 (0.187 cost)	69.50 ± 5.59 (0.412 cost)
NB_{cost}	81.31 ± 2.37 (0.187 cost)	69.52 ± 5.58 (0.398 cost)
Second-year	3 classes	7 classes
NB	71.29 ± 5.68 (0.294 cost)	58.20 ± 6.56 (0.739 cost)
NB_{cost}	71.46 ± 5.72 (0.287 cost)	58.20 ± 6.78 (0.730 cost)
Third-year	4 classes	12 classes
NB	54.26 ± 6.38 (0.488 cost)	50.96 ± 6.63 (1.685 cost)
NB_{cost}	55.02 ± 6.94 (0.481 cost)	51.02 ± 6.49 (1.645 cost)
Fourth-year	4 classes	15 classes
NB	49.65 ± 7.71 (0.540 cost)	50.89 ± 7.38 (4.094 cost)
NB_{cost}	50.21 ± 7.90 (0.539 cost)	49.51 ± 7.35 (4.091 cost)

predictive features (Dataset$_{fs}$). The accuracy and the standard deviations associated with Dataset$_{nofs}$ and Dataset$_{fs}$ are 68.85±5.78 and 69.52±5.58, respectively. Note that most of the models obtain better classification performance when feature selection is performed. Furthermore, it is observed that senior models always use more predictive features than junior models to predict the increase of the *h-index* value within the first four years. Finally, the symbol (†) placed beside a result indicates that the selected model is statistically better than its opposite model (learned from the other dataset) at a specified significance level of 0.05.

In order to determine if the accuracy values are reasonable, Table 11.4 compare cost-sensitive naive Bayes with the standard formulation of naive Bayes. It shows the estimated accuracy, the standard deviations, the average cost (between parentheses) and the number of values of the class variable for each model. Taking the prediction values of junior models for the second year as an example, it is observed that the class variable has 3 possible Δh values (0,1,>1). On the one hand, the accuracy and the standard deviations associated with the naive Bayes and cost-sensitive naive Bayes are 71.29±5.68 and 71.46±5.72, respectively. These accuracy values are considerably greater than would be expected purely by chance. On the other hand, the average cost associated with the naive Bayes and cost-sensitive naive Bayes are 0.294 and 0.287, respectively. Focusing on each algorithm, the cost-sensitive naive Bayes predicts almost all the values more accurately than the naive Bayes. Furthermore, all models obtain lower average cost when the cost-sensitive naive Bayes is used.

Finally, an example of the prediction of the *h-index*'s annual increase is reported. Table 11.5 shows the parameters that define the model associated with senior professors in the first year of prediction. The continuous features are described by means of the mean (μ) and the standard deviation (σ). On the other hand, the discrete feature is described by means of the probability of each possible feature value given the class value ($p(x_i|c)$). The Laplace estimator is used to compute the parameters of conditional distributions of discrete features. Table 11.5 does not show all the parameters for space reasons. Given a senior professor ($\mathbf{x}$) with the following values: *university*=Granada, *documents*=20, *citations*=65, *g-index*=8, *hg-index*=8.4, *m-index*=10, h_r*-index*=9.2, *c-index*=25.3 and h_c*-index*=1.8, the Δh*-index* values can be predicted using the formulation of cost-sensitive naive Bayes and the parameters listed in Table 11.5. After propagating the above evidence, the results predicted by cost sensitive naive Bayes are $p(\Delta h{=}0|\mathbf{x})$=0.004, $p(\Delta h{=}1|\mathbf{x})$=0.308, $p(\Delta h{=}2|\mathbf{x})$=0.688, $p(\Delta h{=}3|\mathbf{x})$=0.000 and $p(\Delta h{=}4|\mathbf{x})$=0.000, that is, with a high probability, the *h-index* of the above professor will increase by two units in the next first year.

11.3 Discussion and conclusions

Machine learning community is not only interested in maximizing classification accuracy, but also in minimizing the expected total cost error derived from mistakes in the classification process. Some ideas, like the cost-sensitive approach, are proposed to face this problem. The proposed cost-sensitive naive Bayes classifier readjusts the probability thresholds of each class to select the class with the minimum-expected cost. This method has been tested on

the bibliometric indices prediction area.

Considering the popularity of the well-know *h-index*, this chapter is focused on building junior and senior models to predict the *h-index* of professors of Spanish public universities in a four-year time horizon. These models are learnt from author-based variables (area, position, university, seniority) and 12 bibliometric indices. The use of models capable of predicting the *h-index* that a researcher will have in coming years can be a useful tool for the scientific community.

Results show that the proposed cost-sensitive naive Bayes usually achieved higher accuracy values and lower average cost than the original naive Bayes, so the cost-sensitive approach could be also applied in different probabilistic classification methods to improve accuracy and costs. Results also show that it is easier to predict the *h-index* of the one-year time horizon than the others, that is, it has a higher average accuracy and lower average total cost than the others. Similarly, it is easier to predict the *h-index* of junior professors than senior professors. Finally, it is observed that specific values of some bibliometric indices can influence the *h-index* value. The probabilities assigned and the predictive features depend on the specific model and time horizon.

In the future, the target will be to build new models that incorporate other researcher-based features, e.g., the impact factor of their journal papers, their collaboration network, percentage of papers published in international journals, among others. Furthermore, some wrapper feature subset selection and other classification methods, e.g., tree augmented network, k-nearest neighbour, C4.5, among others will be used to predict future *h-index* values.

Table 11.5: Parameters that define cost-sensitive naive Bayes model

Features	Δh=0	Δh=1	Δh=2
university	$p(X_i{=}1\|\Delta h{=}0)$	$p(X_i{=}1\|\Delta h{=}1)$	$p(X_i{=}1\|\Delta h{=}2)$
documents	μ=11.7, σ=11.3	μ=17.8, σ=14.9	μ=25.7, σ=9.9
citations	μ=31.1, σ=51.5	μ=63.3, σ=118.6	μ=132.6, σ=108.7
g-index	μ=4.3, σ=3.5	μ=6.3, σ=4.9	μ=10.2, σ=4.8
hg-index	μ=3.2, σ=2.5	μ=4.5, σ=3.7	μ=7.3, σ=3.2
m-index	μ=6.4, σ=6.1	μ=8.7, σ=6.7	μ=12.2, σ=7.8
h_r*-index*	μ=3.3, σ=2.0	μ=4.2, σ=2.9	μ=6.6, σ=2.1
c-index	μ=5.9, σ=8.9	μ=12.4, σ=29.2	μ=18.0, σ=7.9
h_c*-index*	μ=0.4, σ=0.6	μ=0.8, σ=0.8	μ=1.3, σ=0.9

Chapter 12

Uncovering predictive indicators

12.1 Introduction

Many funding agencies and promotion committees use bibliometric indices regularly as a decision-support tool to evaluate individual researchers according to their production. They constitute an objective method whose results are reproducible and can summarize the scientific production of an author as a set of figures. Many works [73, 165, 208, 239, 257, 285, 461, 462] have turned their attention to the predictive power of bibliometric indices in many situations: prediction of article impact, scientist promotions, and acceptance of grant proposals, among others. The result is that the scientific community now faces the challenge of selecting which of this pool of bibliometric indices have a higher predictive power.

Recent works have turned their attention to the predictive power of bibliometric indices in many situations. Jensen *et al.* [239] found that the *h-index* had the best performance to predict the promotions of CNRS researchers. Cabezas-Clavijo *et al.* [73] showed that the main bibliometric indicators that explain the granting of Spanish research proposals in most cases are the number of documents and the number of documents published in first-quartile journals of the Journal Citations Report. Vieira *et al.* [461, 462] assessed the power of models based on bibliometric indicators for the prediction of the rankings of applicants to an academic position at Portuguese universities. They found that models composed by indicators related with the quantity and impact of scientific production, impact of the publication source, prestige of affiliation institution and collaboration provided good predictions and may help peers in their selection process.

This chapter presents a method for identifying a core set of bibliometric indices for prediction purposes, i.e., relevant indices which have a higher predictive power for forecasting other bibliometric indices. Given a dataset of bibliometric indices $\mathbf{X}=\{X_1, X_2, ..., X_n\}$, the task of selecting which subset of bibliometric indices best corresponds to predictive variables $\boldsymbol{X}_P$ (variables with a higher predictive power) and which group can be considered as response variables $\boldsymbol{X}_R$ is tackled, where $dim(\boldsymbol{X}_P)=p$, $dim(\boldsymbol{X}_R)=r$ and $p+r=n$. The best split of predictive and response variables is unknown beforehand and needs to be investigated. The resulting predictive indices are very useful for prediction purposes, that is, when the

relevant index values (predictive variables) are known, knowledge of any index value provides no information on the prediction of other bibliometric indices (response variables).

A wrapper analysis to evaluate all possible configurations of predictor and response variables is used to reveal the relevant set of predictive bibliometric indices. After setting a specific configuration of predictive and response variables, the statistical relationships among the set of bibliometric indices are learned by means of Gaussian Bayesian networks. The learnt network is then used to identify how bibliometric indices relate to each other multivariately, that is, which index X_i is independent of another X_j or which index X_i is conditionally independent of index X_j given the value of a third index X_k, among others. Taking into account the learnt relationships among bibliometric indices, the response variables are predicted by means of multi-output regression [62]. The goal of multi-output regression is to induce a model to simultaneously predict the response variables using the same set of predictive variables and accounting for the dependencies between them.

The structural learning of the Gaussian Bayesian networks, given fixed values for $\boldsymbol{X}_P$ and $\boldsymbol{X}_R$, is based on a score+search approach. This approach optimizes the learning of the structures based on the distance between real and predicted response variable values. The optimization process searches for the Gaussian Bayesian network which minimizes the above fitness score. However, the number of possible structures is huge, and therefore a genetic algorithm [210] is used to explore the search domain of structures. Finally, the optimal structure provides information on which bibliometric indices have the highest predictive power and how they relate to each other.

The interest and originality of this analysis is a novel multi-output regression problem where the role of each variable (predictor or response) is unknown beforehand. To solve this problem, a new Gaussian Bayesian network structure learning algorithm is proposed. It explores the best structure that minimizes the distance between real and predicted response variable values. The resulting structure reports the most predictive bibliometric indices. From their values, a highly accurate of the values of bibliometric indices could be calculated. The scientific community could take advantage of the highly accurate predictions, avoiding the tedious and time-consuming phases of downloading citation records, organizing the non-structured data and computing many bibliometric index values. This chapter is based on the accepted paper [218].

Chapter outline

Section 12.2 briefly introduces the multi-output regression problem and how it can be learned using Gaussian Bayesian networks. Section 12.3 describes the different elements of the genetic algorithm on which the Gaussian Bayesian network learning process is based. Section 12.4 reports the results of applying the proposed approach to a dataset of Spanish full professors of computer science. It covers the experimental setup and the optimal Gaussian Bayesian networks and a discussion on the best induced Gaussian Bayesian network including its network structures, its conditional (in)dependencies among indices and its predictive power.

Finally, conclusions are discussed in Section 12.5.

12.2 Multi-output regression and Gaussian Bayesian networks

The multi-output regression problem is formally described as follows. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random vectors where $\boldsymbol{X}$ consists of p predictive variables and $\boldsymbol{Y}$ consists of r response variables. Given a set of training samples, the goal in multi-output regression is to learn a model which, given an input vector $\boldsymbol{x}$, is able to predict an output vector $\boldsymbol{y}$ that best approximates (in terms of minimizing the least squared errors) the real output vector. Conventionally, this is achieved by generalizing single output regression, using a different regression coefficients vector to predict each output, i.e.,

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{e}, \tag{12.1}$$

where $\boldsymbol{B}$ is a $p \times r$ matrix of regression coefficients, $\boldsymbol{x}$ is a realization of the p predictive variables, and $\boldsymbol{e}$ is a vector consisting of the noise for each of the r response variables. The noise is typically assumed to be Gaussian with a zero mean and uncorrelated across the r response variables.

The multi-output regression problem can be tackled using a Gaussian Bayesian network framework. This framework introduces an alternative parameterization of the regression model derived as a conditional probability model $(\boldsymbol{Y}|\boldsymbol{X})$ from the joint probability distribution. If in the partition $(\boldsymbol{X}, \boldsymbol{Y})$ $\boldsymbol{X}$ is the set of evidential (observed) variables and $\boldsymbol{Y}$ is the set of non-evidential variables, it is assumed a joint multivariate Gaussian distribution with mean vector and covariance matrix given by

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_X} \\ \boldsymbol{\mu_Y} \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{XX}} & \boldsymbol{\Sigma_{XY}} \\ \boldsymbol{\Sigma_{YX}} & \boldsymbol{\Sigma_{YY}} \end{pmatrix}, \tag{12.2}$$

where $\boldsymbol{\mu_X}$ and $\boldsymbol{\Sigma_{XX}}$ are the mean vector and covariance matrix of $\boldsymbol{X}$, $\boldsymbol{\mu_Y}$ and $\boldsymbol{\Sigma_{YY}}$ are the mean vector and covariance matrix of $\boldsymbol{Y}$, and $\boldsymbol{\Sigma_{XY}} = (\boldsymbol{\Sigma_{YX}})^T$ is the covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

The above covariance matrix, $\boldsymbol{\Sigma}$, is of great interest in Gaussian Bayesian networks because its inverse matrix, the precision matrix ($\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$), captures the dependence structure of the variables of the problem. Anderson [21] demonstrated that a variable X_i is conditionally independent of a variable X_j given the rest of the variables iff the value w_{ij}=0. Previously, Shachter and Kenley [411] found that it is possible to determine the precision matrix $\boldsymbol{W}$ using the following recursive formula:

$$\boldsymbol{W}(i+1) = \begin{pmatrix} \boldsymbol{W}(i) + \dfrac{\boldsymbol{\beta}_{i+1}\boldsymbol{\beta}_{i+1}^T}{v_{i+1}} & \dfrac{-\boldsymbol{\beta}_{i+1}}{v_{i+1}} \\ \dfrac{-\boldsymbol{\beta}_{i+1}^T}{v_{i+1}} & \dfrac{1}{v_{i+1}} \end{pmatrix}, \tag{12.3}$$

where $\boldsymbol{W}(i)$ denoted the $i \times i$ upper-left submatrix of $\boldsymbol{W}$, $\boldsymbol{\beta}_{i+1}$ is the i-dimensional vector of

coefficients $\{\beta_{ij}|j<i\}$, and $\boldsymbol{W}(1)=1/v_1$.

Evidence propagation refers to the process of computing the probability distribution of the rest of variables given some observations. A method [81] is used to perform evidence propagation in a Gaussian Bayesian network. Given the above joint distribution, the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is multivariate Gaussian with mean vector $\boldsymbol{\mu}^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$ and covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$ given by

$$\boldsymbol{\mu}^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}} = \boldsymbol{\mu}_{\boldsymbol{Y}} + \boldsymbol{\Sigma}_{\boldsymbol{YX}}\boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}}) \ , \tag{12.4}$$

$$\boldsymbol{\Sigma}^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{YY}} - \boldsymbol{\Sigma}_{\boldsymbol{YX}}\boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{XY}} \ . \tag{12.5}$$

Finally, note that Equations (12.1) and (12.4-12.5) are different parameterizations of the same regression model, since $\boldsymbol{B} = \boldsymbol{\Sigma}_{\boldsymbol{YX}}\boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-1}$.

12.3 Learning Gaussian Bayesian networks using genetic algorithms

Genetic algorithms are stochastic search methods employed in solving complex optimization problems. They mimic the biological mechanisms of natural selection and evolution by means of a fitness function which determines the ability of an individual to survive and reproduce. Genetic algorithms try to find better individuals (solutions for the given problem) by producing fitter descendants in a set of populations.

Genetic algorithms and Gaussian Bayesian networks are used in this chapter to uncover the subset of bibliometric indices with the highest predictive power of all. A wrapper analysis to evaluate all different structures is used to accomplish this goal. Therefore, after setting up a specific splitting of predictive and response variables, a genetic algorithm is used to search the optimal Gaussian Bayesian network structure which minimizes the distance between real and predicted response variable values. The process is repeated for all possible configurations of predictor and response nodes. Figure 12.1 shows the genetic algorithm methodology.

Details of the implementation, such as individual codification, fitness function, selection, crossover, mutation and termination criterion follow.

12.3.1 Initial population

The search space of candidate solutions is represented as a collection of N individuals, called population. In this problem, individuals represent Gaussian Bayesian network structures. Each structure is described by an adjacency matrix $Adj(G)$, which is the representation of the graph $G = (V(G), A(G))$. The adjacency matrix is an $n \times n$ matrix with entries a_{ij}, $i, j = 1, \ldots, n$, such that $a_{ij} = 1$ if and only if an arc exists between nodes i and j, and $a_{ij} = 0$ otherwise. Using this codification, an individual can be transformed into a binary

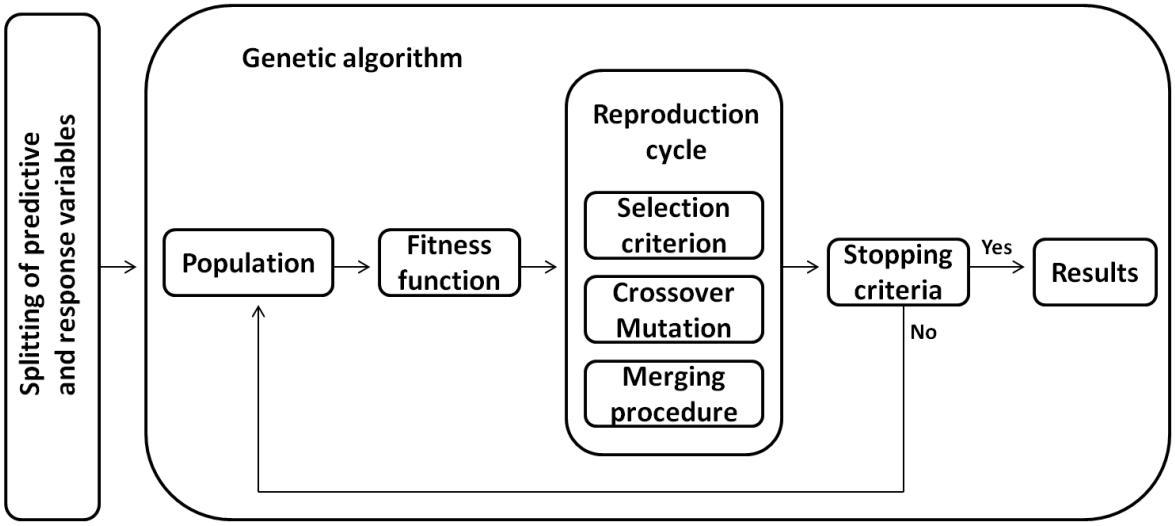


Figure 12.1: Steps of our genetic algorithm methodology.

string (a_{11},...,a_{1n},a_{21},...,a_{2n},...,a_{n1},...,a_{nn}) which maps its adjacency matrix in a vectorized form.

The initial population is randomly generated. Arcs in the adjacency matrices are randomly drawn from a Bernoulli distribution with p=0.5 (probability of sucess). If need be, the network structure is amended to avoid the presence of cycles.

12.3.2 Fitness function

The calculation of the fitness function accounts for the main computational burden in a genetic algorithm. An ideal fitness function should correlate closely with the goal and should be computed quickly. This running time is crucial since the genetic algorithm must be iterated several times to produce reliable results in nontrivial problems.

In this study, given an individual with p predictive variables and r response variables, the Mahalanobis distances between real and predicted values for the r response variables (see Equation (12.6)) are calculated. The Mahalanobis distance is then used as the fitness score of that individual. Considering that the aim is to minimize that distance, the lower the fitness score, the fitter an individual is. The Mahalanobis distance between two vectors $\boldsymbol{y}$ and $\boldsymbol{y}^{'}$ is defined as

$$MD(\boldsymbol{y},\boldsymbol{y}^{'}) = \sqrt{(\boldsymbol{y}-\boldsymbol{y}^{'})^T \Sigma_{\boldsymbol{YY}}^{-1} (\boldsymbol{y}-\boldsymbol{y}^{'})}, \tag{12.6}$$

where $\boldsymbol{y}$ and $\boldsymbol{y}^{'}$ are vectors representing the real and predicted values of the response variables and $\Sigma_{\boldsymbol{YY}}$ is the covariance matrix of the response variables.

The Mahalanobis distance is used as a novel fitness score instead of usual metrics such as K2, BIC and AIC, among others. Although this is a time-consuming fitness value to use (because the predicted values have to be calculated beforehand), a distance-based score is selected because the objective is to minimize the distance between real and predicted response variable values. The Mahalanobis distance is use instead of the Euclidean distance because it

takes into consideration the correlations between all response variables using the covariance matrix, and it solves the problems of scale inherent in the Euclidean distance.

12.3.3 Reproduction cycle

Parent selection criterion, crossover and mutation operators and merging procedure are the constituents of the reproduction cycle in a genetic algorithm. The details of each part are explored below.

Selection criterion The selection process determines which of the individuals from the current population will mate to create new individuals. In general, the fittest individuals will have higher probability of being selected as parents of the next population. Different strategies, like proportional selection methods, ranking selection, tournament selection, etc., are available in the literature [416]. An elitist strategy which chooses the best k individuals from the population as parents for reproduction is used. This strategy guarantees the improvement of the average and minimum value in each genetic algorithm iteration. Thus, the best $N/2$ individuals from the population for reproduction are identified and then moved into a mating pool where they are combined by crossover and mutation operations.

Crossover and mutation Within crossover, $N/2$ parents are randomly mated in pairs to create $N/2$ new children by combining their genotypic information. The aim of crossover is to produce fitter individuals by exchanging information contained in already good individuals [428]. The single-point crossover operator is selected. It is the most common operator and provides good results [250]. Given the binary codification of the individuals, a crossover point is randomly chosen, with a fixed probability P_c, at which the information is exchanged. Based on this point, the strings of both parents are split into two segments each. The first offspring takes the first section from the first parent and the last part from the second, whereas the second offspring is formed conversely.

The mutation operator introduces some extra variability into the population to enhance the diversity degree. It operates on each of the individuals output by crossover by producing random changes with a very small probability. These changes may in turn result in new individuals with higher fitness scores. Single-point mutation is used in this analysis, that is, a bit from the binary string is chosen and flipped with a mutation probability P_m. Finally, whenever an offspring violates the directed acyclic graph constraint, the operator randomly deletes some arcs to amend cycles.

Merging procedure The last stage of the reproductive cycle is the generation of the new population. Again, an elitist strategy is chosen to yield the new population of individuals by combining the best individuals of the previous and new generations. The main advantage of this strategy is that it always preserves the best subset of individuals in every generation.

12.3.4 Stopping criteria

The search is halted when a set of conditions, the stopping criteria, are satisfied. Different criteria for stopping a genetic algorithm have been developed in the literature: after a specific number of generations or a maximum number of evaluations, if there is no improvement in the objective function, or when the objective function outputs a specific value, among others. Here, a maximum number of generations or no improvement over a given number of generations constituted the stopping criteria. The individual with the highest score in the final population is considered to be the solution to the optimization problem.

12.4 Resulting Gaussian Bayesian networks

Different bibliometric indices are used as features variables in the learning Gaussian Bayesian network process. These measures (*documents*, *citations*, *h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, q^2*-index*, h_r*-index*, h_i*-index*, h_c*-index*, and *c-index*) are associated with Spanish full professors of computer science who were active as of January 1^{st}, 2010. This small subset of well-know indices is selected to provide a practical example using the proposed methodology. The selected indices are very popular bibliometric indicators for assessing individual scientists and have an influence on bibliometric and scientometric research. Despite this, they are not the best indices for the above purpose since most of them are size-dependent indicators which sometimes behaves in a counterintuitive manner [309, 468]. In this way, there are better bibliometric indicators like highly cited publications indicators, percentile-based indicators, field-normalized indicators, journal based indicators or collaboration indicators, among others, to evaluate the research performance of scientists. The objective is not to argue in favor of the selected indicators as good ones to assess scientists, they are selected as variables of the Gaussian Bayesian network models. Given these variables, the goal is to simultaneously predict a set of response variables from a set of predictive variables where the role of each variable (predictor or response) is unknown beforehand.

Table 12.1: Statistical figures of all bibliometric indices computed from the publications dataset of Spanish full professors of computer science (years 1973 to 2010).

Variables	Min	1st-quartile	Mean	Median	3rd-quartile	Max
X_1 *(documents)*	1.0	11.3	34.8	21.5	27.2	178.0
X_2 *(citations)*	1.0	17.0	143.0	50.5	145.1	4,570.0
X_3 *(h-index)*	1.0	2.0	6.0	4.0	4.8	37.0
X_4 *(g-index)*	1.0	4.0	11.0	7.0	8.8	66.0
X_5 *(hg-index)*	1.0	2.8	8.4	5.3	6.5	49.4
X_6 *(a-index)*	1.0	5.5	17.8	10.0	14.1	97.5
X_7 *(m-index)*	1.0	5.0	14.0	8.0	11.5	73.0
X_8 *(q^2-index)*	1.0	3.2	9.4	5.5	7.2	52.0
X_9 *(h_r-index)*	1.7	3.0	7.0	9.4	5.8	38.0
X_{10} *(h_i-index)*	0.1	0.6	1.9	1.1	1.5	12.7
X_{11} *(h_c-index)*	0.0	0.0	2.0	1.0	1.3	10.0
X_{12} *(c-index)*	0.3	4.4	19.7	9.2	26.4	908.4

Table 12.2: Correlation coefficients among bibliometric indices computed from the publications dataset of Spanish full professor.

Vars	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
X_1	1.00	0.65	0.78	0.74	0.77	0.52	0.45	0.68	0.78	0.65	0.69	0.61
X_2	-	1.00	0.87	0.88	0.88	0.77	0.70	0.86	0.87	0.83	0.81	0.97
X_3	-	-	1.00	0.96	0.99	0.78	0.71	0.94	0.99	0.93	0.89	0.82
X_4	-	-	-	1.00	0.99	0.86	0.78	0.96	0.96	0.90	0.91	0.83
X_5	-	-	-	-	1.00	0.83	0.75	0.96	0.99	0.92	0.91	0.83
X_6	-	-	-	-	-	1.00	0.94	0.90	0.78	0.74	0.84	0.74
X_7	-	-	-	-	-	-	1.00	0.88	0.71	0.67	0.78	0.65
X_8	-	-	-	-	-	-	-	1.00	0.94	0.88	0.91	0.81
X_9	-	-	-	-	-	-	-	-	1.00	0.93	0.89	0.81
X_{10}	-	-	-	-	-	-	-	-	-	1.00	0.80	0.82
X_{11}	-	-	-	-	-	-	-	-	-	-	1.00	0.75
X_{12}	-	-	-	-	-	-	-	-	-	-	-	1.00

X_1 (*documents*), X_2 (*citations*), X_3 (*h-index*), X_4 (*g-index*), X_5 (*hg-index*), X_6 (*a-index*), X_7 (*m-index*), X_8 (q^2*-index*), X_9 (h_r*-index*), X_{10} (h_i*-index*), X_{11} (h_c*-index*), X_{12} (*c-index*)

In order to give an overview of indices values, Table 12.1 shows a statistical summary for each index. Note that the average academic publishes 34.8 *documents* and receives 143.0 *citations*. Also noticeable is that the average *h-index* value is 6. Interestingly, *citations* values range from 1 to 4,570 citations during the period, i.e., there is at least one academic who has been cited only once, whereas other academics have received a much higher number of citations. The mean citations value (143.0) is on the right of the median value (50.5), which means that the distribution is skewed to the right. This effect is apparent for almost all the indices (except the h_r*-index*). The explanation for this shift is that very few academics excel in terms of productivity, visibility, individuality, innovation and contemporariness.

Table 12.2 shows the correlation coefficients among bibliometric indices. Most of the selected bibliometric indices are variants, extensions or generalizations of the well-known *h-index*. This implies that these indices are usually correlated among them which is corroborated by the mean correlation coefficient (ρ=0.82) among selected indices. Note that *documents* is the index which has the lowest correlations, that is, there are weak correlations between *documents* and *m-index* (ρ=0.45), *documents* and *a-index* (ρ=0.52), *documents* and *c-index* (ρ=0.61), among others. In contrast, the *hg-index* has the highest correlations, that is, there are strong correlations between *hg-index* and *h-index* (ρ=0.99), *hg-index* and *g-index* (ρ=0.99), *hg-index* and h_r*-index* (ρ=0.99), among others.

12.4.1 Experimental setup

The application of a genetic algorithm means setting several parameters like the population size, probabilities for crossover and mutation or the number of allowed iterations. Its efficiency is thus dependent on the chosen parameters. Although some researchers calculate *ad hoc* settings for their specific problem, there are general suggestions which work consistently well for function optimization [117, 188]. In this chapter, Grefenstette's recommendations are

followed with minor changes.

According to Grefenstette [188], the population size should be of 30 individuals. Here, the number of individuals is reduced to 20 because the fitness function is time-consuming to compute. Even so, the population size is enough to uncover the subset of bibliometric indices with the highest predictive power. Crossover probability is 0.9, whereas mutation probability is set at 0.01. A single-point coupled crossover operator and a single-point mutation operator are used in this algorithm. The algorithm halts after reaching 40 generations or when there is no improvement after five consecutive generations.

All different structures with p predictive variables and r response variables are explored. Since the dataset includes 12 bibliometric indices, there are a total of 4,096 ($=2^{12}$) different splittings. Once the role of predictive and response nodes has been fixed, the genetic algorithm searches for the optimal network structure which minimizes the distance between the real and predicted values. The average Mahalanobis distance is used as the fitness function of each individual. Finally, in order to have a fair performance estimation, k-fold cross-validation is chosen as the procedure for estimating the predictive accuracy.

12.4.2 Optimal Gaussian Bayesian networks

The result of the genetic algorithm is a set of 11 optimal Gaussian Bayesian networks. Each model is associated with a different cardinality of predictive and response variables, that is, 1 predictive variable and 11 response variables, 2 predictive variables and 10 response variables and so on. To assess the improvement produced by each optimal network, two general and specific baseline values are firstly defined for comparison. Both of these baselines correspond to the Mahalanobis distances between predicted and real values of the response variables when naive network structures are considered, i.e., networks without arcs between nodes.

The general baseline corresponds to the average Mahalanobis distance of all naive structures with the same number of response variables, regardless of which variables they are. Conversely, the specific baseline accounted for the Mahalanobis distance of a naive structure using the same response variables as the network used for comparison. The rationale behind this is to confirm that the optimal Gaussian Bayesian networks are better than both general and specific baselines.

Table 12.3 shows the list of predictive variables of the 11 optimal Gaussian Bayesian models, general and specific baseline values and the fitness score for each of them. Note that particular fitness scores improve baseline values in all cases. Taking the model with two predictors as an example, it is observed that the predictive variables are X_3 (*h-index*) and X_4 (*g-index*). The set of response variables are X_1 (*documents*), X_2 (*citations*), X_5 (*hg-index*), X_6 (*a-index*), X_7 (*m-index*), X_8 (q^2*-index*), X_9 (h_r*-index*), X_{10} (h_i*-index*), X_{11} (h_c*-index*) and X_{12} (*c-index*). Its associated fitness (2.680) is lower than both baselines (3.013 and 2.931), that is, the predictions of the identified network clearly outperform a naive model with the same number of response variables (3.013) and with the same splitting of variables (2.931).

It is also of interest to compare the performance of the optimal Gaussian Bayesian net-

works with each other. To do so, the fitness of the models is computed given the data. Classical *goodness-of-fit* criteria rank complex models higher than sparse ones. Nonetheless, a model should only have enough parameters to give an adequate representation of the association structure underlying the data. A criterion accounting for this tradeoff between model complexity and goodness-of-fit is the Bayesian information criterion which provides a quantitative measure for model selection. The model with the highest Bayesian information criterion value is selected as the best induced model. Table 12.4 collects the Bayesian information criterion score of each optimal Gaussian Bayesian network. The highest Bayesian information criterion value of all (-6574.755) is achieved by the network with four predictive variables (X_2 (*citations*), X_4 (*g-index*), X_8 (q^2-*index*) and X_9 (h_r-*index*)). Section 12.4.3 details its full structure, conditional dependencies and predictive performance.

12.4.3 Best induced Gaussian Bayesian network

The network which performs best within its class (networks with four predictive variables) and also across the board, is composed of X_2 (*citations*), X_4 (*g-index*), X_8 (q^2-*index*) and X_9 (h_r-*index*) as predictive variables.

Network structure Figure 12.2 illustrates the network structure using blue circles for the predictive variables, and red circles for the response variables. Blue arcs correspond to arcs between predictive variables, whereas red arcs correspond to arcs between response variables. Finally, arcs from predictive to response variables are in black.

A set of centrality measures is examined in order to analyze the graphical characteristics of the network in Figure 12.2. Centrality degree is defined as the number of arcs incident upon a node. Degree is often interpreted in terms of the opportunity for influencing any other node. Two separate measures of centrality degree are defined: indegree and outdegree. A node's indegree is the number of arcs directed to the node, and outdegree is

Table 12.3: Predictive variables for the identified optimal Gaussian Bayesian networks: general and specific baselines and fitness value for the reported model.

Number of predictors	Optimal predictive variables within each network	General baseline	Specific baseline	Best fitness
1	X_9	3.202	3.145	3.025
2	X_3, X_4	3.013	2.931	2.680
3	X_3, X_4, X_{11}	2.817	2.812	2.287
4	X_2, X_4, X_8, X_9	2.612	2.605	1.941
5	X_2, X_5, X_6, X_9, X_{12}	2.398	2.335	1.580
6	X_1, X_2, X_5, X_8, X_9, X_{12}	2.173	2.120	1.228
7	X_2, X_4, X_6, X_8, X_9, X_{10}, X_{12}	1.934	1.905	0.835
8	X_1, X_2, X_3, X_5, X_6, X_7, X_{10}, X_{12}	1.675	1.642	0.381
9	X_1, X_2, X_3, X_5, X_6, X_7, X_{10}, X_{11}, X_{12}	1.389	1.377	0.248
10	X_1, X_2, X_3, X_5, X_6, X_7, X_8, X_{10}, X_{11}, X_{12}	1.058	1.111	0.111
11	X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_{10}, X_{11}, X_{12}	0.649	0.681	0.006

X_1 (*documents*), X_2 (*citations*), X_3 (*h-index*), X_4 (*g-index*), X_5 (*hg-index*), X_6 (*a-index*), X_7 (*m-index*), X_8 (q^2-*index*), X_9 (h_r-*index*), X_{10} (h_i-*index*), X_{11} (h_c-*index*), X_{12} (*c-index*)

Table 12.4: Predictive variables for the identified optimal Gaussian Bayesian networks: Bayesian information criterion values for the reported model.

No. predictors	Predictive variables	BIC values
1	X_9	-7783.949
2	X_3, X_4	-7028.680
3	X_3, X_4, X_{11}	-7020.569
4	X_2, X_4, X_8, X_9	-6574.755
5	X_2, X_5, X_6, X_9, X_{12}	-7106.992
6	X_1, X_2, X_5, X_8, X_9, X_{12}	-6816.705
7	X_2, X_4, X_6, X_8, X_9, X_{10}, X_{12}	-6871.926
8	X_1, X_2, X_3, X_5, X_6, X_7, X_{10}, X_{12}	-6933.521
9	X_1, X_2, X_3, X_5, X_6, X_7, X_{10}, X_{11}, X_{12}	-6655.728
10	X_1, X_2, X_3, X_5, X_6, X_7, X_8, X_{10}, X_{11}, X_{12}	-7232.618
11	X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_{10}, X_{11}, X_{12}	-8037.581

X_1 (*documents*), X_2 (*citations*), X_3 (*h-index*), X_4 (*g-index*), X_5 (*hg-index*), X_6 (*a-index*), X_7 (*m-index*), X_8 (q^2*-index*), X_9 (h_r*-index*), X_{10} (h_i*-index*), X_{11} (h_c*-index*), X_{12} (*c-index*)

the number of arcs that the node directs to others. Therefore, indegree is the number of parents, whereas outdegree is the number of children. The centrality degree (CD) values are: CD(X_1)=5, CD(X_2)=5, CD(X_3)=7, CD(X_4)=10, CD(X_5)=8, CD(X_6)=6, CD(X_7)=7, CD(X_8)=7, CD(X_9)=8, CD(X_{10})=6, CD(X_{11})=7 and CD(X_{12})=6. Note that the *g-index* (X_4) has a great influence on other indices: it presents the highest centrality degree (2 + 8 = 10). Also worth mentioning is that indices such as h_r*-index* (X_9) and h_i*-index* (X_{10}) show opposite structures, that is, h_r*-index* (X_9) has no parents but eight children, whereas h_i*-index* (X_{10}) depends on six parents but has no children. At the other end of the scale, *documents* (X_1) and *citations* (X_2) show the lowest centrality degree with a value of 5, suggesting that they have very little influence on the other indices.

Focusing on potential relationship between strong correlation coefficients and network structure, we observe that strong correlations are not a problem for the methodology presented. A potential high correlation coefficient does not imply an arc among correlated variables in the Gaussian Bayesian network. In this way, Table 12.2 shows strong correlations between *hg-index* and h_r*-index* (ρ=0.99) and between *h-index* and h_i*-index* (ρ=0.93) which are not presented as arcs in Figure 12.2. In contrast, the weak correlation between *documents* and *m-index* (ρ=0.45) is presented as an arc in Figure 12.2. The presence of arcs does not depend on potential strong correlations, it depends on the genetic algorithm which looks for the optimal structure that minimize the distance between real and predicted response variable values.

Dependencies among indices Based on the definitions of the indices, it is clear that some of them can be expressed according to the values of other indices. For example, *hg-index* could be expressed in terms of *h-* and *g-index* values. Also, q^2*-index* can be defined according to *h-* and *m-index* values. This is corroborated by the dependencies in the network. The *h-index* (X_3) and the *g-index* (X_4) are parent nodes of *hg-index* (X_5) in the network structure of Figure 12.2, and *h-index* (X_3) and *m-index* (X_7) are children of the q^2*-index*

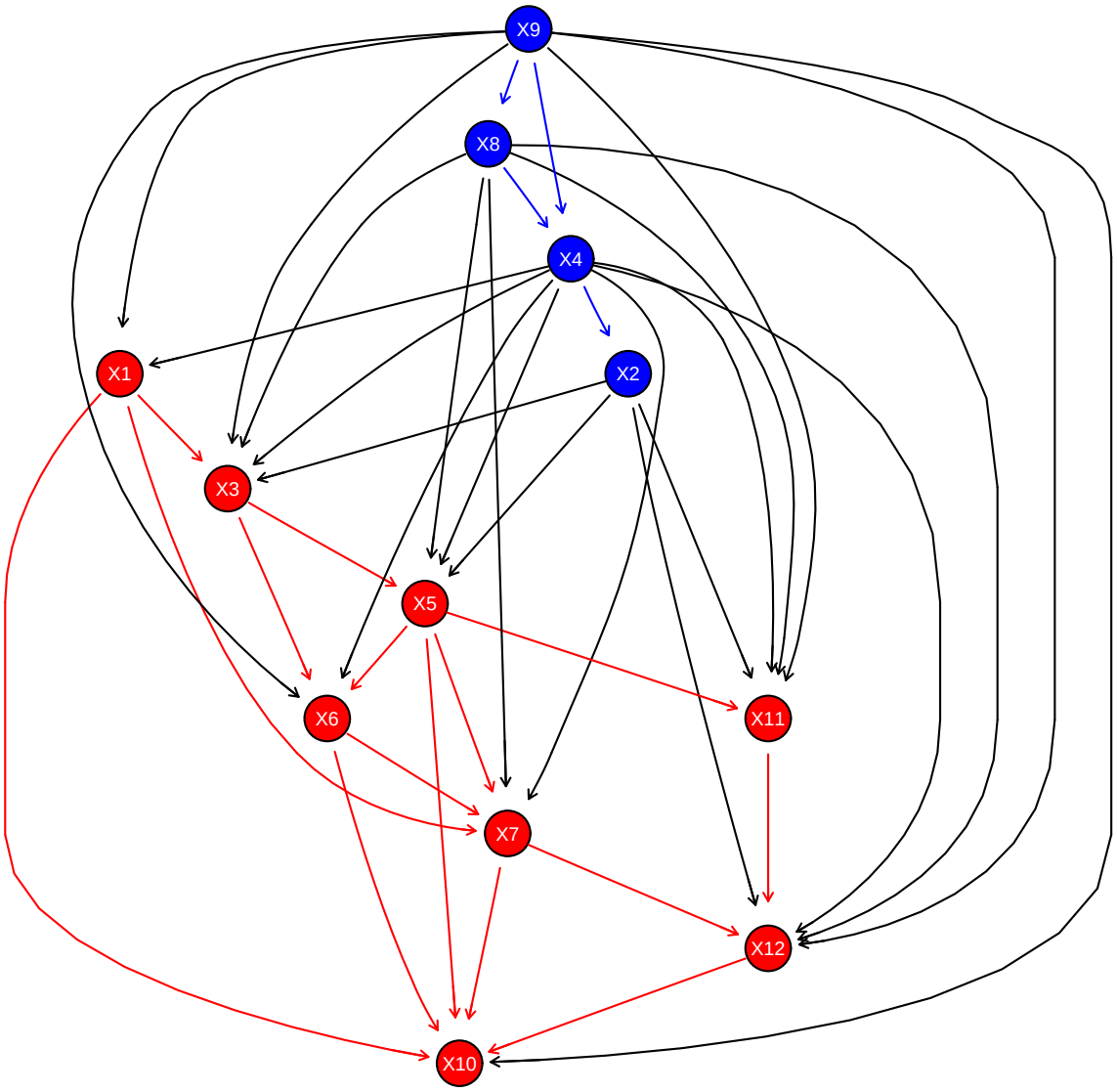


Figure 12.2: Best GBN structure. Each node represents: X_1 (*documents*), X_2 (*citations*), X_3 (*h-index*), X_4 (*g-index*), X_5 (*hg-index*), X_6 (*a-index*), X_7 (*m-index*), X_8 (q^2*-index*), X_9 (h_r*-index*), X_{10} (h_i*-index*), X_{11} (h_c*-index*), X_{12} (*c-index*). Blue nodes correspond to predictive variables ($\boldsymbol{X_P}$), whereas red nodes correspond to response variables ($\boldsymbol{X_R}$).

(X_8).

Besides unveiling dependencies already present in the index definitions, the best Gaussian Bayesian network discovers dependencies which are related to but not directly derived from, but related to index definitions. Taking the arc from h_r*-index* to *h-index* ($X_9 \rightarrow X_3$) as an example, note that the information about h_r*-index* influences the density function of *h-index*, as expected in h_r*-index*, an extension of *h-index*.

The arc between *a-index* and *m-index*, ($X_6 \rightarrow X_7$) in Figure 12.2 is an example of a dependency that is initially expected. Remember that *a-index* represents the average number

of citations received by the articles included in the *h-core*, whereas the *m-index* represents the median number of citations received by the articles in the same *h-core*. Therefore, both refer to citations of articles in the *h-core*.

Other dependencies, such as the arcs between *documents* and *h-index* ($X_1 \rightarrow X_3$), *citations* and *h-index* ($X_2 \rightarrow X_3$), *g-index* and *citations* ($X_4 \rightarrow X_2$), or *g-index* and *h-index* ($X_4 \rightarrow X_3$), are not immediately apparent from the definitions. Nevertheless, they have been reported to show a high value of correlation [57, 101, 396]. There are other network dependencies, e.g., *g-index* and h_c*-index* ($X_4 \rightarrow X_{11}$), and h_c*-index* and h_i*-index* ($X_{11} \rightarrow X_{10}$), which cannot be linked to the individual definitions. However, previous works have already pointed out similar correlations [152].

Conversely, the network included some unexpected arcs. In this way, the Gaussian Bayesian network reported probabilistic dependencies between *a-index* and h_i*-index* ($X_6 \rightarrow X_{10}$), *m-index* and h_c*-index* ($X_7 \rightarrow X_{11}$), and q^2*-index* and *c-index* ($X_8 \rightarrow X_{12}$).

Conditional independencies among indices Gaussian Bayesian networks are a powerful tool not only for capturing dependencies but also for identifying conditional independencies among variables. Here, Markov network properties are used within the aim of discovering such independencies among the nodes of the best induced network. The *local Markov property* states that any node X_i is conditionally independent of its *non-descendants* given the values of its *parents*. It can be expressed as $I(X_i, \text{non-descendants}(X_i) \mid \Pi(X_i))$. With respect to a whole network, the *global Markov property* states that any node X_i is conditionally independent of any other node given the values of its Markov blanket (MB). The Markov blanket of a node includes its parents, its children, and its children's parents. Thus, $I(X_i, \text{non-}MB(X_i) \mid MB(X_i))$.

Table 12.5 lists conditional independencies between the bibliometric indices of the network in Figure 12.2. The list is derived from the local and global Markov properties. This network identifies conditional independencies in accordance with index definitions, as well as other conditional independencies that are hidden in such definitions. Taking *hg-index* as an example, it is observed that, given *citations*, *h-index*, *g-index* and q^2*-index*, *hg-index* is independent of *documents* and h_r*-index*. This suggests that when the values of *citations*, *h-index*, *g-index* and q^2*-index* are known, the value of *documents* provides no information on the value of *hg-index*. Focusing on q^2*-index*, note that it is conditionally independent of h_i*-index* given its MB, which includes *h-index* and *m-index*, among others. Some other reasonable conditional independency relationships are also listed in Table 12.5.

However, there are other conditional independencies which are not obvious. According to the definition of the *a-index*, it is reasonable to expect that it is dependent on *documents* and *citations*. Nevertheless, the model shows that *a-index* is conditionally independent of *documents* and *citations*, given *h-index*, *g-index*, *hg-index* and h_r*-index*. Similarly, one might expect a dependency relationship between *citations* and *documents*, but the model suggests that the relationship is of conditional independency given *g-index*. Remember that the conditional independencies between indices encoded in the Gaussian Bayesian network indicate

Table 12.5: Conditional independencies among bibliometric indices derived using local and global Markov properties in the Gaussian Bayesian network of Figure 12.2.

Index	is conditionally independent of	given
documents	*citations*, q^2-*index*	*g-index*, h_r-*index*
documents	h_c-*index*	*citations*, *h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, q^2-*index*, h_r-*index*, h_i-*index*, *c-index*
citations	*documents*, q^2-*index*, h_r-*index*	*g-index*
citations	*a-index*, h_i-*index*	*documents*, *h-index*, *g-index*, *hg-index*, *m-index*, q^2-*index*, h_r-*index*, h_c-*index*, *c-index*
h-index	*m-index*, h_i-*index*, h_c-*index*, *c-index*	*documents*, *citations*, *g-index*, *hg-index*, *a-index*, q^2-*index*, h_r-*index*
g-index	h_i-*index*	*documents*, *citations*, *h-index*, *hg-index*, *a-index*, *m-index*, q^2-*index*, h_r-*index*, h_c-*index*, *c-index*
hg-index	*documents*, h_r-*index*	*citations*, *h-index*, *g-index*, q^2-*index*
a-index	*documents*, *citations*, q^2-*index*, h_c-*index*	*h-index*, *g-index*, *hg-index*, h_r-*index*
a-index	*citations*, h_c-*index*	*documents*, *h-index*, *g-index*, *hg-index*, *m-index*, q^2-*index*, h_r-*index*, h_i-*index*, *c-index*
m-index	*citations*, *h-index*, h_r-*index*, h_c-*index*	*documents*, *g-index*, *hg-index*, *a-index*, q^2-*index*
m-index	*h-index*	*documents*, *citations*, *g-index*, *hg-index*, *a-index*, q^2-*index*,, h_r-*index*, h_i-*index*, h_c-*index*, *c-index*
q^2-*index*	h_i-*index*	*documents*, *citations*, *h-index*, *g-index*, *hg-index*, *a-index*, *m-index*, h_r-*index*, h_c-*index*, *c-index*
h_i-*index*	*citations*, *h-index*, *g-index*, q^2-*index*, h_c-*index*	*documents*, *hg-index*, *a-index*, *m-index*, h_r-*index*, *c-index*
h_c-*index*	*documents*, *h-index*, *a-index*, *m-index*	*citations*, *g-index*, *hg-index*, q^2-*index*, h_r-*index*
h_c-*index*	*documents*, *h-index*, *a-index*, h_i-*index*	*citations*, *g-index*, *hg-index*, *m-index* q^2-*index*, h_r-*index*, *c-index*
c-index	*documents*, *h-index*, *hg-index*, *a-index*	*citations*, *g-index*, *m-index*, q^2-*index*, h_r-*index*, h_c-*index*
c-index	*h-index*	*documents*, *citations*, *g-index*, *hg-index*, *a-index**m-index*, q^2-*index*, h_r-*index*, h_i-*index*, h_c-*index*

a probabilistic not a causal relationship.

Predicting bibliometric indices Now, the probabilistic component of the network in Figure 12.2 and what the effect of knowing the values of some variables has on the others are inspected. In doing so, evidence propagation is used to compute the probability distribution of other variables given the available evidence. Using the values of the predictive variables, that is, *citations* (X_2), *g-index* (X_4), q^2-*index* (X_8) and h_r-*index* (X_9), the Gaussian Bayesian network is able to predict the (expected) values of the response variables: *documents* (X_1), *h-index* (X_3), *hg-index* (X_5), *a-index* (X_6), *m-index* (X_7), h_i-*index* (X_{10}), h_c-*index* (X_{11}), and *c-index* (X_{12}).

Table 12.6 presents three inference examples. It shows the evidence values for the predictive variables and the predictions made by the network in Figure 12.2 for the response variables. These predictions are the mean vector ($\boldsymbol{\mu}^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$) of the conditional distribution of

Table 12.6: Evidence propagation results using the Gaussian Bayesian network of Figure 12.2 as the inference tool.

Variables	Example 1		Example 2		Example 3	
Predictive	Evidences		Evidences		Evidences	
citations	853.0		163.0		10.0	
g-index	28.0		12.0		3.0	
q^2*-index*	19.4		10.2		2.4	
h_r*-index*	15.9		8.9		2.8	
Responses	Predicted	Real	Predicted	Real	Predicted	Real
documents	74.8	73.0	44.4	43.0	14.1	13.0
h-index	14.9	15.0	8.0	8.0	1.9	2.0
hg-index	20.4	20.5	9.8	9.8	2.4	2.4
a-index	43.5	42.9	14.4	14.0	4.1	3.0
m-index	24.4	25.0	12.6	13.0	3.6	3.0
h_i*-index*	5.1	3.9	2.7	1.6	0.5	0.4
h_c*-index*	4.0	5.0	1.8	1.0	0.4	0.0
c-index	78.9	80.6	15.4	14.8	2.4	2.8

$\boldsymbol{Y}$ given $\boldsymbol{X}$, which is computed with Equation (12.4). The real values of the response variables are also shown for comparison against predictions. Three different examples ranging from high, medium and low values are set as evidence.

In Example 1, it is fixed *citations*=853, *g-index*=28, q^2*-index*=19.4 and h_r*-index*=15.9. After setting the evidence, the predicted values of the response indices were computed. Results in Table 12.6 show that predicted values are very close to real values. Regarding *documents*, *h-index* and *hg-index*, it is observed that predictions are 74.8, 14.9 and 20.4, whereas the real values were 73.0, 15.0 and 20.5, respectively. In Example 2, values of *citations*=163, *g-index*=12, q^2*-index*=10.2 and h_r*-index*=8.9 are set as evidences. Predictions are again very close to actual values. Remarkably, predicted and real values are equal for *h-index* and *hg-index*. Lastly, Example 3 sets *citations*=10, *g-index*=3, q^2*-index*=2.4 and h_r*-index*=2.8. Given these values, differences between real and predicted values are also slight.

12.5 Discussion and conclusions

Bibliometric indices are an increasingly important topic for the scientific community nowadays. Many bibliometric indices have been developed in order to consider previously uncovered aspects. In this context, some researchers have recently turned their attention to the predictive power of bibliometric indices in many situations. The result is that the scientific community now faces the challenge of selecting which of this pool of bibliometric indices have a higher predictive power.

A review of the literature presents some recent works [461, 462] which analyzed how good are models based on bibliometric indices in predicting the rankings of applicants to academic positions at the university. Like this work, they learned different models to assess the predictive power of bibliometric indices. Their rank ordered logistic regression models were composed by indicators related to the quantity and impact of scientific production,

impact of the publication source, prestige of affiliation institution and collaboration. Unlike this multi-output regression approach, they only predict a response variable. Also, they did not face the problem of selecting the role (predictor or response) of each variable. Finally, their results suggested that the models could predict the result of peer-review with a reasonable degree of accuracy.

Unlike the above studies, a novel method is presented to uncover a relevant core subset of indicators for prediction purposes given a set of bibliometric indices. The selected bibliometric indices are popular indicators to evaluate individual scientists and also have an influence on the scientific community. Despite this, most of them are size-dependent indicators which sometimes behaves in a counterintuitive way because of the inconsistencies associated with the mechanism used to aggregate publication and citation statistics into a single number. This work does not argue in favor of the selected indicators as the best bibliometric indices to evaluate research performance, they are selected as an example of Gaussian Bayesian network variables in order to give a practical example using the proposed methodology.

Given a dataset of bibliometric indices, it is tackled the task of selecting which subset of bibliometric indices best correspond to predictive variables and which group can be considered as response variables. The goal is to simultaneously predict a set of response variables from a set of predictive variables by means of multi-output regression. This results in a novel multi-output regression problem where the role of each variable is unknown beforehand.

Fixed a specific splitting of predictive and response variables, a Gaussian Bayesian network structure is learned to identify relationships among bibliometric indices and for prediction purposes. Gaussian Bayesian network structure learning is based on a genetic algorithm which optimizes the distance between real and predicted response variable values. The best network is the one that minimizes the Mahalanobis distance between real and predicted values, and has the highest Bayesian information criterion value among the 11 optimal models with different cardinality. Although an exhaustive analysis is used to evaluate all possible configurations of predictive and response variables in order to identify the relevant predictive core set of bibliometric indices, a genetic algorithm could be also used for exploring the search domain of different configurations of predictive and response variables.

In this specific problem with full professors, the findings provide information on which subset of bibliometric indices has the highest predictive power, i.e., is more relevant for prediction purposes. Note that the bibliometric core is composed of *citations*, *g-index*, q^2-*index* and h_r-*index*. This means that when the values of the above bibliometric indices are known, the values of the other eight indices can be predicted with high accuracy. Analyzing its structure, it is observed that it matches many expected dependencies among indices. In addition, the model is able to discover new knowledge when combined with the index definitions and sheds light on unreported conditional (in)dependencies between the indices.

Finally, the proposed methodology does not require any specific values of predictive and response variables. Also, it is not affected by specifications such as the number of observations or variables of the dataset. In this way, the methodology can be applied to any dataset. Obviously, the methodology results usually depend on the selected dataset. Despite this,

similar bibliometric indices relationships could be also learnt using different bibliometric datasets.

In the future, alternative models will be learn using different Bayesian network induction algorithms. It would also be worthwhile to extend the domain of the data collection to overseas all Spanish researchers. These new models could also incorporate other bibliometric indices in order to cover a larger part of the bibliometric domain.

Part V

Conclusions

Chapter 13

Conclusions and Future work

13.1 Summary of contributions

The main and specific conclusions drawn from this thesis have been presented throughout each chapter. The most relevant will be summarized in this chapter, emphasizing the reached achievements. These contributions have been divided into three parts:

Predicting bibliometric indices: Based on the motivation that publishers of scientific journals face the tough task of selecting high quality articles that will attract as many citations as possible from a pool of articles, Chapter 5 presents predictive models that forecast the citation count of journal articles within first few years after publication. Results show that the logistic regression and naive Bayes classification methods output high average scores in the different journal sections and across the time horizon. These models also found that the appearance of certain words in the paper abstracts can influence the number of citations received. These selected tokens could be used as a point of reference to identify the hot topics. Finally, the use of these models capable of predicting the citations that an article will receive in the first few years after publication can be a useful tool for publishers' assessment process, paving the way for new assessment systems.

The scientific community now focuses on the popular *h-index* since it is used by funding agencies and promotion committees to evaluate the importance of research. In this context, Chapter 6 and Chapter 11 presents cost-sensitive models to predict the annual increase of the *h-index*. Specially, Chapter 6 forecast the annual increase for journals using cost-sensitive selective naive Bayes models that use directly the misclassification costs in the learning algorithms. In contrast, Chapter 11 forecast the annual increase for researchers using cost-sensitive naive Bayes models that take into account the expected cost of instances predictions at classification time. Results show that proposed models outperform many cost-(in)sensitive models, so this learning approach could be used in different probabilistic classification approaches.

Discovering new associations among indices: The vast number of existing bibliometric indices poses the challenge of exploiting the relationships among them. Based on this challenge, Chapter 7 and Chapter 12 analyze the current relationships and discover new conditional (in)dependencies between bibliometric indices using Bayesian networks. Specially, Chapter 7 analyzes bibliometric indices on computer science and artificial intelligence journals, whereas Chapter 12 analyzes bibliometric indices of Spanish full professors associated with the computer science area. Using the proposed models, scientific community could measure how indices influence others in probabilistic terms and perform evidence propagation and abduction inference for answering bibliometric questions.

Besides the above goals, Chapter 12 presents a new method to uncover which bibliometric indices have a higher predictive power. It tackles the task of selecting which subset of bibliometric indices best correspond to predictive variables and which group can be considered as response variables. This method solves a proposed multi-output regression problem where the role of each variable is unknown beforehand. Gaussian Bayesian networks and genetic algorithms are used to achieve the predictive core set of bibliometric indices. Results show that the optimal induced Gaussian Bayesian networks corroborate previous relationships between several indices but also suggest new previously unreported interactions. Also, results show that a set of 12 bibliometric indices can be accurately predicted using only a smaller predictive core subset composed of four indices. These four indices are very useful for prediction purposes, that is, when their index values are known, the rest of indices can be accurately predicted.

Exploring Spanish computer science research: Based on the publish or perish culture, researchers' behavior has been affected in the sense that it is not only important what researchers write, but also how often, where and with whom they write. In this context, Chapter 8, Chapter 9 and Chapter 10 analyze how this culture affects Spanish computer science research.

A scientometric analysis of the Spanish computer science research is achieved in Chapter 8. Results show that Spanish research productivity and visibility have increased their values in the last years, achieving an increment of 347% and 1,053%, respectively. Spanish academics usually publish more proceeding papers than journal articles despite of the low number of citations received by proceeding papers. They also publish more documents in high quality journals than previous years. Regarding universities, some universities such as Universidad de Granada, Universidad Politécnica de Catalunya, Universidad Politécnica de Valencia, Universidad Politécnica de Madrid and Universidad de Málaga usually rank top positions for different parameters. Finally, by subject areas, computer languages and systems academics publish the highest number of Spanish computer science documents, whereas computer science and artificial intelligence academics excel in terms of citation per document, documents per academic, citations per academic and percentage of documents published in high quality journals.

The productivity, visibility, quality, prestige and international collaboration of Spanish computer science research is also analyzed using a cluster analysis methodology in Chapter 9. Results of the proposed methodology show that Spanish public universities fall into four different clusters, whereas academic staff belong into six different clusters. For example, universities like Universidad de Granada, Universidad de Jaén, Universidad Pablo de Olavide de Sevilla and Universidad Pública de Navarra score highest for productivity, visibility and quality. Universities like Universidad del País Vasco and Universidad Politécnica de Madrid excel in terms of prestige, whereas universities like Universidad Politécnica de Valéncia stand out on international collaboration. The resulting clusters could have potential implications on research policy, proposing collaborations and alliances among universities, supporting institutions in the processes of strategic planning, and verifying the effectiveness of research policies, among others.

Finally, Chapter 10 studied the number of documents and citations by number of authors. These measures are also analyzed when documents are written in different types of collaboration (international, national and institutional), when documents are published in different document types (journal article and conference paper), when documents are published in different computer science subdisciplines (artificial intelligence, cybernetics, hardware and architecture, information systems, interdisciplinary applications, software engineering and theory and methods), and, finally, when documents are published by journals with different impact factor quartiles (first-quartile journals, second-quartile journals, third-quartile journals and fourth-quartile journals). Results did not find a positive association between author set cardinality and the citation impact.

13.2 List of publications

The publications derived from this research [218, 219, 220, 221, 222, 223, 224, 225, 226] are listed below

- Ibáñez, A. Armañanzas, R., Bielza, C. and Larrañaga, P. (2015). Genetic algorithms and Gaussian Bayesian networks to uncover the predictive core set of bibliometric indices. *Journal of the Association for Information Science and Technology*, accepted. Current Impact factor: 2.230. Ranking: Q1 (17/135). Category: Computer Science, Information Systems.

- Ibáñez, A., Bielza, C. and Larrañaga, P. (2014). Cost-sensitive selective naive Bayes classifiers for predicting the increase of the h-index for scientific journals. *Neurocomputing*, 135:42-52. Current Impact factor: 2.005. Ranking: Q1 (28/121). Category: Computer Science, Artificial Intelligence. http://dx.doi.org/10.1016/j.neucom.2013.08.042

- Ibáñez, A., Larrañaga, P. and Bielza, C. (2013). Cluster methods for assessing research performance: exploring Spanish computer science. *Scientometrics*, 97(3):571-600. Current Impact factor: 2.274. Ranking: Q1 (20/102). Category: Computer Science, Interdisciplinary Applications. http://dx.doi.org/10.1007/s11192-013-0985-9

- Ibáñez, A., Bielza, C. and Larrañaga, P. (2013). Relationship among research collaboration, number of documents and number of citations. A case study in Spanish computer science production in 2000-2009. *Scientometrics*, 95(2):689-716. Current Impact factor: 2.274. Ranking: Q1 (20/102). Category: Computer Science, Interdisciplinary Applications. http://dx.doi.org/10.1007/s11192-012-0883-6

- Ibáñez, A., Bielza, C. and Larrañaga, P. (2013). Analysis of scientific activity in Spanish public universities in the area of computer science. *Revista Española de Documentación Científica*, 36(1):e002. Current Impact factor: 0.717. Ranking: Q2 (40/83). Category: Information Science and Library Science. http://dx.doi.org/10.3989/redc.2013.1.912

- Ibáñez, A., Larrañaga, P. and Bielza, C. (2011). Predicting the h-index with cost-sensitive naive Bayes. *Proceedings of the 11th International Conference on Intelligent Systems Design and Applications*, 599-604, http://dx.doi.org/10.1109/ISDA.2011.6121721

- Ibáñez, A., Bielza, C. and Larrañaga, P. (2011). Productividad y Visibilidad Científica de los Profesores Funcionarios de las Universidades Públicas Españolas en el Área de Tecnologías Informáticas. *Fundación General de la UPM*, ISBN: 978-84-15302-05-6. http://oa.upm.es/9407/

- Ibáñez, A., Larrañaga, P. and Bielza, C. (2011). Using Bayesian network to discover relationships between bibliometric indices. A case study of computer science and artificial intelligence journals. *Scientometrics*, 89(2):523-551. Current Impact factor: 2.274. Ranking: Q1 (20/102). Category: Computer Science, Interdisciplinary Applications. http://dx.doi.org/10.1007/s11192-011-0486-7

- Ibáñez, A., Larrañaga, P. and Bielza, C. (2009). Predicting citation count of Bioinformatics papers within four years of publication. *Bioinformatics*, 25(24):3303-3309. Current Impact factor: 4.621. Ranking: Q1 (4/52). Category: Mathematical and Computational Biology. http://dx.doi.org/10.1093/bioinformatics/btp585

13.3 Future work

Several interesting problems related to this research are still open for future investigation. This section summarizes and emphasizes the most relevant future lines and open issues that have been already enumerated through the specific conclusion section of each chapter.

According to the prediction of bibliometric indices, this dissertation proposes predictive models that forecast the citation count of journal articles within first few years after publication using tokens found in the abstracts of papers. The target will be to build new predictive models that incorporate other paper-based features (title, keywords, conclusions, etc.), new author-based features (*h-index*, number of papers, number of citations, etc.) and new journal-based features (impact factor, immediacy index, category, etc.) as predictive variables. These models would be induced using different machine learning methods like regression, regularized regression, or local regression which model the number of citations as a

continuous variable. This research also presents cost-sensitive models to predict the annual increase of the *h-index*. The aim will be to build new cost-sensitive Bayesian classifiers like the selective tree augmented naive Bayes which could perform better in terms of accuracy and cost. Beside matrices of misclassification cost, other methods which express relative distances between classes could be also considered. Furthermore, other interesting bibliometric indices like collaboration-based measures (percentage of documents published by single authors, percentage of documents published in international collaboration and so on) could be used as predictive features to forecast the scientific success of journals and researchers.

By discovering new associations among bibliometric indices, this thesis analyzes the relationships among well-known measures using Bayesian networks. Besides the proposed models, alternative Bayesian network models will be induced using different learning algorithms based on both constraint-based methods and score+search methods. Other probabilistic graphical models could also be taken into consideration. These new models will incorporate other indices that take time into account, allow for co-authorship, assess the quality of citations, and correct for differences among fields, among others, in order to cover a larger part of the bibliometric domain. Regarding uncovering which bibliometric indices have a higher predictive power, the proposed method exhaustively evaluate all possible configurations of predictive and response variables in order to identify the relevant predictive core set of bibliometric indices. In this context, an optimization algorithm will be used, instead of an exhaustive analysis, for exploring the search domain of different configurations of predictive and response variables

Further updated scientometrics analysis will be carried out to characterize the Spanish computer science research activity. The target will be to incorporate private universities and non-tenured academics. Also, other aspects (number of patents, number of projects, number of spin-offs, number of different co-authors, proximity among co-authors, among other) will be taken into account. Regarding the effect of research collaboration, it will be analyzed whether researchers with the best research performance are also the investigators that collaborate more at the international level, and whether the citation counts of papers that have been written by authors with a low number of citations improve through collaboration. Other scientific disciplines will be also analyzed to get a comprehensive overview of the Spanish research. Finally, the bibliometric indices values vary depending on the source consulted (Web of Science, Scopus and Google Scholar). It is a point to be taken into account in future research.

To conclude, it is important to remark that the proposed machine learning methods could be applied to any dataset. In this context, not only the scientometrics discipline is benefited, but also the rest of scientific disciplines could use the novel techniques proposed in this dissertation.

Bibliography

[1] http://thomsonreuters.com/thomson-reuters-web-of-science/.

[2] http://www.scopus.com/.

[3] http://scholar.google.com/.

[4] S. Abe. *Support Vector Machines for Pattern Classification*. Springer, 2010.

[5] G. Abramo and C. A. D'Angelo. National-scale research performance assessment at the individual level. *Scientometrics*, 86(2):347–364, 2011.

[6] G. Abramo, C. A. D'Angelo, and F. Pugini. The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology. *Scientometrics*, 76(2):225–244, 2008.

[7] G. Abramo, C.A. D'Angelo, and M. Solazzi. Are researchers that collaborate more at the international level top performers? An investigation on the Italian university system. *Journal of Informetrics*, 5(1):204–213, 2011.

[8] G. Abramo, C.A. D'Angelo, and M. Solazzi. The relationship between scientists' research performance and the degree of internationalization of their research. *Scientometrics*, 86(3):629–643, 2011.

[9] D. E. Acuna, S. Allesina, and K. P. Kording. Future impact: Predicting scientific success. *Nature*, 489:201–202, 2012.

[10] R. Adler, J. Ewing, and P. Taylor. Citation statistics. *Statistical Sciences*, 24(1):1–14, 2009.

[11] L. S. Adriaanse and C. Rensleigh. Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. *The Electronic Library*, 31(6):727–744, 2013.

[12] N. Agrait and A. Poves. Report on CNEAI assessment results. Technical report, National Evaluation Committee of Research Activity, 2009. In Spanish.

[13] D. W. Aha. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, 36(2):267–287, 1992.

[14] D. W. Aha. *Lazy Learning*. Springer, 1997.

[15] I. Ajiferuke and D. Wolfram. Citer analysis as a measure of research impact: Library and information science as a case study. *Scientometrics*, 83(3):623–638, 2010.

[16] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[17] S. Alonso, F. J. Cabrerizo, E. Herrera-Viedma, and F. Herrera. h-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3(4):273–289, 2009.

[18] S. Alonso, F. J. Cabrerizo, E. Herrera-Viedma, and F. Herrera. hg-index: A new index to characterize the scientific output of researchers based on the h- and g-indices. *Scientometrics*, 82(2):391–400, 2010.

[19] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2004.

[20] T. R. Anderson, R. K. S. Hankin, and P. D. Killworth. Beyond the Durfee square: Enhancing the h-index to score total publication output. *Scientometrics*, 76(3):577–588, 2008.

[21] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, 2003.

[22] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[23] D. Archibugi and A. Coco. International partnerships for knowledge in business and academia: A comparison between Europe and the USA. *Technovation*, 24(7):517–528, 2004.

[24] N. Assimakis and M. Adam. A new author's productivity index: p-index. *Scientometrics*, 85(2):415–427, 2010.

[25] C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11(1-5):11–73, 1997.

[26] N. Bakkalbasi, K. Bauer, J. Glover, and L. Wang. Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, 3(7):1–8, 2006.

[27] P. Balaram. Scientometrics: A dismal science. *Current Science*, 95(4):431–432, 2008.

[28] P. Ball. Index aims for fair ranking of scientists. *Nature*, 436:900, 2005.

[29] G. Bammer. Enhancing research collaborations: Three key management challenges. *Research Policy*, 37(5):875–887, 2008.

[30] M. Banks. An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics*, 69(1):161–168, 2006.

[31] J. Bar-Ilan. Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2):257–271, 2008.

[32] J. Bar-Ilan. Citations to the "Introduction to informetrics" indexed by WoS, Scopus and Google Scholar. *Scientometrics*, 82(3):495–506, 2010.

[33] J. Bar-Ilan, M. Levene, and A. Lin. Some measures for comparing citation databases. *Journal of Informetrics*, 1(1):26–34, 2007.

[34] C. Bartneck and J. Hu. The fruits of collaboration in a multidisciplinary field. *Scientometrics*, 85(1):41–52, 2010.

[35] T. Bartol, G. Budimir, D. Dekleva-Smrekar, M. Pusnik, and P. Juznic. Assessment of research fields in Scopus and Web of Science from the viewpoint of national research evaluation in Slovenia. *Scientometrics*, 98(2):1491–1504, 2014.

[36] P. D. Batista, M. G. Campiteli, O. Kinouchi, and A. S. Martinez. Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1):179–189, 2006.

[37] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1764.

[38] D. Beaver. Reflections on scientific collaboration (and its study): Past, present and future. *Scientometrics*, 52(3):365–377, 2001.

[39] D. Beaver and R. Rosen. Studies in scientific collaboration. Part I. The professional origins of scientific co-authorship. *Scientometrics*, 1(1):65–84, 1979.

[40] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[41] S. J. Bensman and L. Leydesdorff. Definition and identification of journals as bibliographic and subject entities: Librarianship versus ISI journal citation reports methods and their effect on citation measures. *Journal of the American Society for Information Science and Technology*, 60(6):1097–1117, 2009.

[42] C. Bergstrom. Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, 68(5):314–316, 2007.

[43] C. Bielza and P. Larrañaga. Discrete Bayesian network classifiers: A survey. *ACM Computing Surveys*, 47(1):Article 5, 2014.

[44] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford university Press, 1995.

[45] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[46] R. Blanco, I. Inza, and P. Larrañaga. Learning Bayesian networks in the space of structures by estimation of distribution algorithms. *International Journal of Intelligent Systems*, 18(2):205–220, 2003.

[47] J. Bollen, M. A. Rodriguez, and H. van de Sompel. Journal status. *Scientometrics*, 69(3):669–687, 2006.

[48] J. Bollen, H. van de Sompel, A. Hagberg, and R. Chute. A principal component analysis of 39 scientific impact measures. *Plos One*, 4(6):e6022, 2009.

[49] M. Bordons, R. Sancho, F. Morillo, and I. Gómez. Perfil de actividad científica de las universidades españolas en cuatro áreas temáticas: Un enfoque multifactorial. *Revista Española de Documentación Científica*, 33(1):9–33, 2010.

[50] L. Bornmann and H. D. Daniel. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2006.

[51] L. Bornmann and H. D. Daniel. What do we know about the h-index? *Journal of the American Society for Information Science and Technology*, 58(9):1381–1385, 2007.

[52] L. Bornmann and H. D. Daniel. What do citation counts measure? *Journal of Documentation*, 64(1):45–80, 2008.

[53] L. Bornmann and H. D. Daniel. The citation speed index: A useful bibliometric indicator to add to the h-index. *Journal of Informetrics*, 4(3):444–446, 2010.

[54] L. Bornmann and L. Leydesdorff. Which are the best performing regions in information science in terms of highly cited papers? Some improvements of our previous mapping approaches. *Journal of Informetrics*, 6(2):336–345, 2012.

[55] L. Bornmann, R. Mutz, and H. Daniel. Are there better indices for evaluation purposes than the h-index? A comparison of nine different variants of the h-index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5):830–837, 2008.

[56] L. Bornmann, R. Mutz, and H. D. Daniel. The b-index as a measure of scientific excellence. A promising supplement to the h-index. *International Journal of Scientometrics, Informetrics and Bibliometrics*, 11(1):6, 2007.

[57] L. Bornmann, G. Wallon, and A. Ledin. Is the h-index related to (standard) measures and to the assessments by peers? An investigation of the h-index by using molecular life sciences data. *Research Evaluation*, 17(2):149–156, 2008.

[58] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

[59] T. Braun and E. Bujdoso. Some tendencies of the radioanalytical literature. Statistical games for trend evaluation. *Radiochemical and Radioanalytical Letters*, 23(4):195–203, 1975.

[60] T. Braun, W. Glanzel, and A. Schubert. A Hirsch-type index for journals. *Scientometrics*, 69(1):169–173, 2006.

[61] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. Chapman and Hall, 1993.

[62] L. Breiman and J. H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(1):3–54, 1997.

[63] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[64] T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Association for Information Science and Technology*, 57(8):1060–1072, 2006.

[65] B. C. Brookes. Biblio-, sciento-, infor-metrics??? What are we talking about? In *Selection of Papers Submitted for the Second International Conference on Bibliometrics, Scientometrics and Informetrics*, pages 31–43, 1990.

[66] B. Y. Brusilovsky. Partial and system forecasts in scientometrics. *Technological Forecasting and Social Change*, 12(2-3):193–200, 1978.

[67] P. Buhlmann, M. Kalisch, and M. H. Maathuis. Variable selection in high-dimensional linear models: Partially faithful distributions and the PC-simple algorithm. *Biometrika*, 97(2):261–278, 2010.

[68] A. T. Bui and C. H. Jun. Learning Bayesian network structure using Markov blanket decomposition. *Pattern Recognition Letters*, 33(16):2134–2140, 2012.

[69] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[70] Q. L. Burrell. Hirsch index or Hirsch rate?. Some thoughts arising from Liang's data. *Scientometrics*, 73(1):19–28, 2007.

[71] H. D. Burton. Use of a virtual information system for bibliometric analysis. *Information Processing and Management*, 24(1):39–44, 1988.

[72] G. Cabanac. Shaping the landscape of research in information systems from the perspective of editorial boards: A scientometric study of 77 leading journals. *Journal of the American Society for Information Science and Technology*, 63(5):977–996, 2012.

[73] A. Cabezas-Clavijo, N. Robinson-García, M. Escabias, and E. Jiménez-Contreras. Reviewers' ratings and bibliometric indicators: Hand in hand when assessing over research proposals? *PLoS ONE*, 8(6):e68258, 2013.

[74] F. J. Cabrerizo, S. Alonso, E. Herrera-Viedma, and F. Herrera. q^2-index: Quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core. *Journal of Informetrics*, 4(1):23–28, 2010.

[75] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[76] J. M. Campanario, W. Cabos, and M. A. Hidalgo. El impacto de la producción científica de la Universidad de Alcalá de Henares. *Revista Española de Documentación Científica*, 21(4):402–415, 1998.

[77] F. B. F. Campbell. *The Theory of the National and International Bibliography: with Special Reference to the Introduction of System in the Record of Modern Literature.* Library Bureau, 1896.

[78] A. Cano, M. Gómez-Olmedo, and S. Moral. Approximate inference in Bayesian networks using binary probability trees. *International Journal of Approximate Reasoning*, 52(1):49–62, 2011.

[79] J. S. Cardodo and J. F. Pinta da Costa. Learning to classify ordinal data: The data replication method. *The Journal of Machine Learning Research*, 8:1393–1429, 2007.

[80] C. Castillo, D. Donato, and A. Gionis. Estimating the number of citations using author reputation. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, volume 4726, pages 107–117, Santiago, Chile, 2007. Springer.

[81] E. Castillo, J. M. Gutierrez, and A. S. Hadi. *Expert Systems and Probabilistic Network Models.* Springer-Verlag, New York, 1997.

[82] J. C. Chai, P. H. Hua, R. Rousseau, and J. K. Wan. The adapted pure h-index. In *Proceedings of the Fourth International Conference on Webometrics, Informetrics and Scientometrics*, pages 1–6, 2008.

[83] X. Chai, L. Deng, Q. Yang, and C. X. Ling. Test-cost sensitive naive Bayes classification. In *Fourth IEEE International Conference on Data Mining*, pages 51–58, 2004.

[84] O. Chapelle, B. Scholkopf, and A. Zien. *Semi-Supervised Learning.* MIT Press, 2006.

[85] P. Cheeseman and J. Stutz. *Bayesian Classification (Autoclass): Theory and Results.* AAAI Press, 1996.

[86] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007.

[87] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. R. Liu. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1-2):43–90, 2002.

[88] S. Cheng, P. YunTao, Z. YanNing, M. Zheng, Y. JunPeng, G. Hong, Y. ZhengLu, M. CaiFeng, and W. YiShan. PrestigeRank: A new evaluation method for papers and journals. *Journal of Informetrics*, 5(1):1–13, 2011.

[89] D. M. Chickering. Learning Bayesian networks is NP-complete. Technical Report MSR-TR-94-17, Microsoft Research, 1994.

[90] D. M. Chickering, D. Geiger, and D. Heckerman. Learning Bayesian networks: Search methods and experimental results. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 112–128, 1995.

[91] G. Chirici. Assessing the scientific productivity of Italian forest researchers using the Web of Science, Scopus and Scimago databases. *iForest - Biogeosciences and Forestry*, 5(3):101–107, 2012.

[92] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[93] E. Cobo, A. Selva-O'Callagham, J. Ribera, F. Cardellach, R. Dominguez, and M. Vilardell. Statistical reviewers improve reporting in biomedical articles: A randomized trial. *PLoS ONE*, 2(3):e332, 2007.

[94] F. J. Cole. The history of comparative anatomy. Part I. A statistical analysis of the literature. *Science Progress*, 11(44):578–596, 1917.

[95] J. R. Cole and S. Cole. *Social Stratification in Science*. University of Chicago Press, 1973.

[96] S. Cole, J. R. Cole, and G. A. Simon. Chance and consensus in peer review. *Science*, 214:881–886, 1981.

[97] G. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2-3):393–405, 1990.

[98] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[99] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*, chapter Greedy algorithms. MIT Press, 1990.

[100] R. Costas and M. Bordons. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, 1(3):193–203, 2007.

[101] R. Costas and M. Bordons. Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, 77(2):267–288, 2008.

[102] R. Costas, T. N. van Leeuwen, and M. Bordons. A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology*, 61(8):1564–1581, 2010.

[103] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[104] J. Cowie, L. Oteniya, and R. Coles. Particle swarm optimization for learning Bayesian networks. In *Proceedings of the World Congress on Engineering*, pages 2–4, 2007.

[105] K. Crammer and Y. Singer. Pranking with ranking. In *Advances in Neural Information Processing Systems*, volume 14, pages 641–647. MIT Press, 2002.

[106] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[107] B. Cronin. The need for a theory of citing. *Journal of Documentation*, 37(1):16–24, 1981.

[108] E. Csajbok, A. Berhidi, L. Vasas, and A. Schubert. Hirsch-index for countries based on Essential Science Indicators data. *Scientometrics*, 73(1):91–117, 2007.

[109] P. W. Cullen, R. H. Norris, V. H. Resh, T. B. Reynoldson, D. M. Rosenberg, and M. T. Barbour. Collaboration in scientific research: A critical need for freshwater ecology. *Freshwater Biology*, 42(1):131–142, 1999.

[110] A. Darwiche. A differential approach to inference in Bayesian networks. *Journal of the ACM*, 50(3):280–305, 2003.

[111] A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.

[112] B. V. Dasarathy. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, 1991.

[113] D. L. Davies and D. W. Bouldin. A clustering separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.

[114] P. M. Davis. Eigenfactor: Does the principle of repeated improvement result in better journal impact estimates than raw citation counts? *Journal of the American Society for Information Science and Technology*, 59(13):2186–2188, 2008.

[115] L. De Campos. Independency relationships and learning algorithms for singly connected networks. *Journal of Experimental and Theoretical Artificial Intelligence*, 10(4):511–549, 1998.

[116] L. De Campos, J. M. Fernández-Luna, J. A. Gámez, and J. M. Puerta. Ant colony optimization for learning Bayesian networks. *International Journal of Approximate Reasoning*, 31(3):291–311, 2002.

[117] K. A. De Jong and W. M. Spears. An analysis of the interacting roles of population size and crossover in genetic algorithms. In *Proceedings of the 1st Workshop on Parallel Problem Solving from Nature*, pages 38–47, 1990.

[118] D. J. De Solla Price. *Science Since Babylon*. Yale University Press, 1961.

[119] D. J. De Solla Price. *Little Science, Big Science*. Columbia University Press, 1963.

[120] D. J. De Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[121] D. J. De Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[122] D. J. De Solla Price. Multiple authorship. *Science*, 212(4498):987, 1981.

[123] D. J. De Solla Price and D. Beaver. Collaboration in an invisible college. *American Psychologist*, 21(11):1011–1018, 1966.

[124] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[125] M. M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, 2009.

[126] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, pages 155–164, 1999.

[127] P. Dong, M. Loh, and A. Mondry. The impact factor revisited. *Biomedical Digital Libraries*, 2(2):7, 2005.

[128] C. Drummond and R. Holte. Exploiting the cost (in)sensitivity of decision tree splitting criteria. In *Proceedings of the 17th International Conference on Machine Learning*, pages 239–246, 2000.

[129] D. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, USA, 1973.

[130] D. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley Interscience, 2001.

[131] A. W. F. Edwards. System to rank scientists was pedalled by Jeffreys. *Nature*, 437:951, 2005.

[132] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of The American Statistical Association*, 78(382):316–331, 1983.

[133] L. Egghe. Dynamic h-index: the Hirsch index in function of time. *Journal of the American Society for Information Science and Technology*, 58(3):452–454, 2006.

[134] L. Egghe. How to improve the h-index. *The Scientist*, 20(3):14, 2006.

[135] L. Egghe. An improvement of the h-index: The g-index. *ISSI Newsletter*, 2(1):8–9, 2006.

[136] L. Egghe. Theory and practice of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[137] L. Egghe. Mathematical study of h-index sequences. *Information Processing and Management*, 45(2):288–297, 2009.

[138] L. Egghe. The Hirsch index and related impact measures. *Annual Review of Information Science and Technology*, 44(1):65–114, 2010.

[139] L. Egghe, R. Rousseau, and G. van Hooydonk. Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the American Society for Information Science*, 51(2):145–157, 2000.

[140] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference of Artificial Intelligence*, pages 973–978, 2001.

[141] Elvira-Consortium. Elvira: An environment for probabilistic graphical models. In *Proceedings of the First European Workshop on Probabilistic Graphical Models (PGM'02)*, pages 222–230, 2002.

[142] R. Etxeberria, P. Larrañaga, and J. M. Pikaza. Analysis of the behaviour of genetic algorithms when learning Bayesian network structure from data. *Pattern Recognition Letters*, 18(11-13):1269–1273, 1997.

[143] B. S. Everitt, S. Landau, and M. Leese. *Cluster Analysis.* Arnold, 2001.

[144] J. Ewing. Measuring journals. *Notices of the American Mathematical Society*, 53(9):1049–1053, 2006.

[145] R. A. Fairthorne. Empirical hyperbolic distributions (bradford-zipf-mandelbrot) for bibliometric description and prediction. *Journal of Documentation*, 25(4):319–343, 1969.

[146] M. E. Falagas and V. G. Alexiou. The top-ten in journal impact factor manipulation. *Archivum Immunologiae et Therapiae Experimentalis*, 56(4):223–226, 2008.

[147] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *FASEB*, 22(2):338–342, 2008.

[148] X. Fang. Inference-based naive Bayes: Turning naive Bayes cost-sensitive. *IEEE Transactions on Knowledge and Data Engineering*, 25(10):2302–2313, 2013.

[149] A. Fersht. The most influential journals: Impact factor and Eigenfactor. *Proceedings of the National Academy of Sciences*, 106(17):6883–6884, 2009.

[150] M. J. Flores, A. E. Nicholson, A. Brunskill, K. B. Korb, and S. Mascaro. Incorporating expert knowledge when learning Bayesian network structure: A medical case study. *Artificial Intelligence in Medicine*, 53(3):181–204, 2011.

[151] L. Fortnow. Time for computer science to grow up. *Communications of the ACM*, 52(8):33–35, 2009.

[152] M. Franceschet. A cluster analysis of scholar and journal bibliometric indicators. *Journal of the American Society for Information Science and Technology*, 60(10):1950–1964, 2009.

[153] M. Franceschet. A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, 83(1):243–258, 2010.

[154] M. Franceschet. The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis. *Journal of Informetrics*, 4(1):55–63, 2010.

[155] M. Franceschet. Journal influence factors. *Journal of Informetrics*, 4(3):239–248, 2010.

[156] M. Franceschet. The role of conference publications in computer science: A bibliometric view. *Communications of the ACM*, 53(12):129–132, 2010.

[157] M. Franceschet. Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, 62(10):1992–2012, 2011.

[158] M. Franceschet and A. Costantini. The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, 4(4):540–553, 2010.

[159] F. Franceschini, M. Galetto, D. Maisano, and L. Mastrogiacomo. The success-index: An alternative approach to the h-index for evaluating an individual's research output. *Scientometrics*, 92(3):621–641, 2012.

[160] F. Franceschini and D. Maisano. Analysis of the Hirsch index's operational properties. *European Journal of Operational Research*, 203(2):494–504, 2010.

[161] E. Frank and M. Hall. A simple approach to ordinal classification. In *Proceedings of the 12th European Conference on Machine Learning*, pages 145–156, 2001.

[162] E. Frank and S. Kramer. Ensembles of nested dichotomies for multi-class problems. In *Proceedings of the 21st International Conference on Machine Learning*, pages 305–312, 2004.

[163] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian networks classifiers. *Machine Learning*, 29:131–163, 1997.

[164] L. D. Fu and C. F. Aliferis. Models for predicting and explaining citation count of biomedical articles. *AMIA Annual Symposium Proceedings*, pages 222–226, 2008.

[165] L. D. Fu and C. F. Aliferis. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, 85(1):257–270, 2010.

[166] J. Furnkranz. Pairwise classification as an ensemble technique. In *Proceedings of the 13th European Conference on Machine Learning*, pages 97–110, 2002.

[167] J. Gama. A cost-sensitive iterative Bayes. In *Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning*, 2000.

[168] M. A. García-Pérez. A multidimensional extension to Hirsch's h-index. *Scientometrics*, 81(3):779–785, 2009.

[169] M. A. García-Pérez. Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h-indices in psychology. *Journal of the American Society for Information Science and Technology*, 61(10):2070–2085, 2010.

[170] E. Garfield. Citation indexes for science. A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, 1955.

[171] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.

[172] E. Garfield. 'Citations to' divided by 'items published' gives journal impact factor. Essays of an information scientist. *Current Contents*, 1(7):270–273, 1972.

[173] E. Garfield. Citation frequency as a measure of research activity and performance. *Essays of an Information Scientist*, 1:406–408, 1973.

[174] E. Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. John Wiley, 1979.

[175] E. Garfield. Use of Journal Citation Reports and Journal Performance Indicators in measuring short and long term journal impact. *Croatian Medical Journal*, 41(4):368–374, 2000.

[176] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.

[177] A. Gazni, C.R. Sugimoto, and F. Didegah. Mapping world scientific collaboration: Authors, institutions, and countries. *Journal of the American Society for Information Science and Technology*, 63(2):323–335, 2012.

[178] D. Geiger and D. Heckerman. Learning Gaussian networks. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 235–243, 1994.

[179] W. Glanzel. National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1):69–115, 2001.

[180] W. Glanzel. On the opportunities and limitations of the h-index. *Science Focus*, 1(1):10–11, 2006.

[181] W. Glanzel and B. Thijs. Does co-authorship inflate the share of self-citations? *Scientometrics*, 61(3):395–404, 2004.

[182] W. Glanzel, B. Thijs, A. Schubert, and K. Debackere. Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, 78(1):165–188, 2009.

[183] I. Gómez Caridad, M.T. Fernández Muñoz, M. Bordons, and F. Morillo. La producción científica española en Medicina en los años 1949-1999. *Revista Clínica Española*, 204(2):75–88, 2004.

[184] R. Golubic, M. Rudes, N. Kovacic, M. Marusic, and A. Marusic. Calculating impact factor: How bibliographical classification of journal items affects the impact factor of large and small journals. *Science and Engineering Ethics*, 14(1):41–49, 2008.

[185] B. González-Albo and M.A. Zulueta García. Patentes domésticas de universidades españolas: Análisis bibliométrico. *Revista Española de Documentación Científica*, 30(1):61–90, 2007.

[186] B. González-Pereira, V. P. Guerrero-Bote, and F. Moya-Anegón. A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, 4(3):379–391, 2010.

[187] A. A. Goodrum, K. W. McCain, S. Lawrence, and C. L. Giles. Scholarly publishing in the internet age: A citation analysis of computer science literature. *Information Processing and Management*, 37(5):661–675, 2001.

[188] J. J. Grefenstette. Optimization of control parameters for genetic algorithms. *IEEE Transactions on Systems, Man and Cybernetics*, 16(1):122–128, 1986.

[189] J. C. Guan and N. Ma. A comparative study of research performance in computer science. *Scientometrics*, 61(3):339–359, 2004.

[190] L. Guerra, V. Robles, C. Bielza, and P. Larrañaga. A comparison of clustering quality indices using outliers and noise. *Intelligent Data Analysis*, 16(4):703–715, 2012.

[191] V. P. Guerrero-Bote and F. Moya-Anegón. A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics*, 6(4):674–688, 2012.

[192] H. Guo and W. Hsu. A survey of algorithms for real-time Bayesian network inference. In *Proceedings of the AAAI Workshop on Real-Time Decision Support and Diagnosis Systems*, pages 1–12, 2002.

[193] G. Haddow and P. Genoni. Australian education journals: Quantitative and qualitative indicators. *Australian Academic and Research Libraries*, 40(2):88–104, 2009.

[194] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2/3):107–145, 2001.

[195] G. Hanks. Peer review in action: The contribution of referees to advancing reliable knowledge. *Palliative Medicine*, 19(5):359–370, 2005.

[196] P. E. Hart. The condensed nearest neighbour rule. *Transactions on Information Theory*, 14:515–516, 1968.

[197] A. W. Harzing, S. Alakangas, and D. Adams. hIa: an individual annual h-index to accommodate disciplinary and career length differences. *Scientometrics*, 99(3):811–821, 2014.

[198] A. W. Harzing and R. van der Wal. Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, 8(1):1–13, 2008.

[199] R. Hauptman. How to be a successful scholar: Publish efficiently. *Journal of Scholarly Publishing*, 36(2):115–119, 2005.

[200] D. T. Hawkins. Unconventional uses of on-line information retrieval systems: On-line bibliometric studies. *Journal of the American Society for Information Science*, 28(1):13–18, 1977.

[201] S. S. Haykin. *Neural Networks and Learning Machines*. Prentice Hall, 2009.

[202] Y. He and J. C. Guan. Contribution of Chinese publications in computer science: A case study on LNCS. *Scientometrics*, 75(3):519–534, 2008.

[203] D. Heckerman. *A Tutorial on Learning with Bayesian Networks*. MIT Press, 1998.

[204] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks. The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[205] M. Henrion. Propagating uncertainty in Bayesian networks by probabilistic logic sampling. In *Proceedings of the Second Annual Conference on Uncertainty in Artificial Intelligence*, pages 149–164, 1986.

[206] D. J. Hess. *Science Studies: An advanced introduction*. New York University Press, 1997.

[207] J. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

[208] J. Hirsch. Does the h-index have predictive power? *Proceedings of the National Academy of Sciences*, 104(49):19193–19198, 2007.

[209] J. Hirsch. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3):741–754, 2010.

[210] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[211] W. W. Hood and C. S. Wilson. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2):291–314, 2001.

[212] D. Horrobin. Something rotten at the core of science. *Trends in Pharmacological Sciences*, 22(2):51–52, 2001.

[213] D. L. Horrobin. The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association*, 263(10):1438–1441, 1990.

[214] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, USA, 2nd edition, 2000.

[215] S. Huang, J. Li, J. Ye, A. Fleisher, K. Chen, T. Wu, E. Reiman, and Alzheimer's Disease Neuroimaging Initiative. A sparse structure learning algorithm for Gaussian Bayesian network identification from high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1328–1342, 2013.

[216] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[217] E. W. Hulme. *Statistical Bibliography in Relation to the Growth of Modern Civilization.* Grafton, 1923.

[218] A. Ibáñez, R. Armañanzas, C. Bielza, and P. Larrañaga. Genetic algorithms and Gaussian Bayesian networks to uncover the predictive core set of bibliometric indices. *Journal of the Association for Information Science and Technology*, Accepted for publication, 2015.

[219] A. Ibáñez, C. Bielza, and P. Larrañaga. Análisis de la actividad científica de las universidades públicas españolas en el área de las tecnologías informáticas. *Revista Española de Documentación Científica*, 36(1):e002, 2013.

[220] A. Ibáñez, C. Bielza, and P. Larrañaga. Relationship among research collaboration, number of documents and number of citations. A case study in Spanish computer science production in 2000-2009. *Scientometrics*, 95(2):689–716, 2013.

[221] A. Ibáñez, C. Bielza, and P. Larrañaga. Cost-sensitive selective naive Bayes classifiers for predicting the increase of the h-index for scientific journals. *Neurocomputing*, 135:42–52, 2014.

[222] A. Ibáñez, C. Bielza, and P.Larrañaga. *Productividad y Visibilidad Científica de los Profesores Funcionarios de las Universidades Públicas Españolas en el Área de Tecnologías Informáticas.* Fundación General de la U.P.M., 2011.

[223] A. Ibáñez, P. Larrañaga, and C. Bielza. Predicting citation count of Bioinformatics papers within four years of publication. *Bioinformatics*, 25(24):3303–3309, 2009.

[224] A. Ibáñez, P. Larrañaga, and C. Bielza. Predicting the h-index with cost-sensitive naive Bayes. In *Proceedings of the 11th International Conference on Intelligent Systems Design and Applications*, pages 599–604, 2011.

[225] A. Ibáñez, P. Larrañaga, and C. Bielza. Using Bayesian networks to discover relationships between bibliometric indices. A case study of computer science and artificial intelligence journals. *Scientometrics*, 89(2):523–551, 2011.

[226] A. Ibáñez, P. Larrañaga, and C. Bielza. Cluster methods for assessing research performance: exploring Spanish computer science. *Scientometrics*, 97(3):571–600, 2013.

[227] J. E. Iglesias and C. Pecharroman. Scaling the h-index for different scientific ISI fields. *Scientometrics*, 73(3):303–320, 2007.

[228] P. Jacso. As we may search - comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89(9):1537–1547, 2005.

[229] P. Jacso. Comparison and analysis of the citedness scores in Web of Science and Google Scholar. In *Proceedings of the International Conference on Asian Digital Libraries*, pages 360–369, 2005.

[230] P. Jacso. The pros and cons of computing the h-index using Google Scholar. *Online Information Review*, 32(3):437–452, 2008.

[231] P. Jacso. The pros and cons of computing the h-index using Scopus. *Online Information Review*, 32(4):524–535, 2008.

[232] P. Jacso. The pros and cons of computing the h-index using Web of Science. *Online Information Review*, 32(5):673–688, 2008.

[233] A. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall College Division, 1988.

[234] A. Jain, M. Murty, and P. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[235] C. Jennings. Citation data: The wrong impact? *Nature Neuroscience*, 1(8):641–642, 1998.

[236] F. Jensen and S. Anderson. Approximations in Bayesian belief universe for knowledge based systems. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 162–169, 1990.

[237] F. V. Jensen. *An Introduction to Bayesian Networks*. UCL Press, 1996.

[238] F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer, 2001.

[239] P. Jensen, J. B. Rouquier, and Y. Croissant. Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*, 78(3):467–479, 2009.

[240] H. Y. Jia, D. Y. Liu, and P. Yu. Learning dynamic Bayesian network with immune evolutionary algorithm. In *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, pages 2934–2938, 2005.

[241] L. Jiang, C. Li, and S. Wang. Cost-sensitive Bayesian network classifiers. *Pattern Recognition Letters*, 45:211–216, 2014.

[242] B. Jin. H-index: An evaluation indicator proposed by scientist. *Science Focus*, 1(1):8–9, 2006.

[243] B. Jin, L. Liang, R. Rousseau, and L. Egghe. The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, 52(6):855–863, 2007.

[244] B. H. Jin. The AR-index: Complementing the h-index. *ISSI Newsletter*, 3(1):6, 2007.

[245] P. Juznic, S. Peclin, M. Zaucer, T. Mandelj, M. Pusnik, and F. Demsar. Scientometric indicators: Peer-review, bibliometric methods and conflict of interests. *Scientometrics*, 85(2):429–441, 2010.

[246] M. Kalisch and P. Buhlmann. Robustification of the PC-algorithm for directed acyclic graphs. *Journal of Computational and Graphical Statistics*, 17(4):773–789, 2008.

[247] D. Katsaros, A. Sidiropoulos, and Y. Manopoulos. Age decaying h-index for social network of citations. In *Proceedings of the BIS 2007 Workshop on Social Aspects of the Web*, page 3, 2007.

[248] J.S. Katz and B.R. Martin. What is research collaboration? *Research Policy*, 26(1):1–18, 1997.

[249] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.

[250] T. Kellegoza, B. Toklub, and J. Wilsonc. Comparing efficiencies of genetic crossover operators for one machine total weighted tardiness problem. *Applied Mathematics and Computation*, 199(2):590–598, 2008.

[251] C. D. Kelly and M. D. Jennions. The h-index and career assessment by numbers. *Trends in Ecology and Evolution*, 21(4):167–170, 2006.

[252] C. D. Kelly and M. D. Jennions. H-index: Age and sex make it unreliable. *Nature*, 449(7161):403, 2007.

[253] I. Khawaja. An alternative stipulation of the term bibliometry. *Pakistan Library Bulletin*, 18(4):1–6, 1987.

[254] A. Khurshid and H. Sahai. Bibliometric distributions and laws: Some comments and a selected bibliography. *Journal of Educational Media and Library Sciences*, 28(4):433–459, 1991.

[255] J. H. Kim and J. Pearl. A computational model for causal and diagnostic reasoning in inference systems. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pages 190–193, 1983.

[256] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. American Mathematical Society, 1980.

[257] I. Kissin. Can a bibliometric indicator predict the success of an analgesic? *Scientometrics*, 86(3):785–795, 2011.

[258] R. Kohavi. *Wrapper for Performance Enhancement and Oblivious Decision Graphs*. PhD thesis, Stanford University, 1995.

[259] D. Koller and N. Friedman. *Probabilistic Graphical Models. Principles and Techniques*. The MIT Press, Massachusetts, 2009.

[260] I. Kononenko. Semi-naive Bayesian classifier. In *Proceedings of the Sixth European Working Session on Learning*, pages 206–219, 1991.

[261] R. Korf. Linear-space best-first search. *Artificial Intelligence*, 62(1):41–78, 1993.

[262] M. Kosmulski. A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter*, 2(3):4–6, 2006.

[263] M. Kosmulski. New seniority-independent Hirsch-type index. *Journal of Informetrics*, 3(4):341–347, 2011.

[264] S. Kotsiantis, I. Zaharakis, and P. Pintelas. Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.

[265] S. B. Kotsiantis. Local ordinal classification. In *Artificial Intelligence Applications and Innovations*, International Federation for Information Processing, pages 1–8. Springer, 2004.

[266] S. B. Kotsiantis and P. E. Pintelas. A cost sensitive technique for ordinal classification problems. In *Methods and Applications of Artificial Intelligence*, Lecture Notes in Computer Science, pages 220–229. Springer, 2004.

[267] S. Kramer, G. Widmer, B. Pfahringer, and M. De Groeve. Prediction of ordinal classes using regression trees. *Fundamenta Informaticae - Intelligent Systems*, 47(1-2):1–13, 2001.

[268] G. Krampen, A. von Eye, and G. Schui. Forecasting trends of development of psychology from a bibliometric perspective. *Scientometrics*, 87(2):687–694, 2011.

[269] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

[270] A. V. Kulkarni, B. Aziz, I. Shams, and J. W. Busse. Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *The Journal of the American Medical Association*, 302(10):1092–1096, 2009.

[271] C. Lacave. *Explanation in Causal Bayesian networks. Medical Applications*. PhD thesis, Dept. Inteligencia Artificial. UNED, Madrid, Spain (in Spanish), 2003.

[272] C. Laine and C.D. Mulrow. Exorcising ghosts and unwelcome guests. *Annals of Internal Medicine*, 143(8):611–612, 2005.

[273] B.S. Lancho Barrantes, V.P. Guerrero Bote, Z. Chinchilla Rodríguez, and F. de Moya Anegón. Citation flows in the zones of influence of scientific collaborations. *Journal of the American Society for Information Science and Technology*, 63(3):481–489, 2012.

[274] R. Landry and N. Amara. The impact of transaction costs on the institutional structuration of collaborative academic research. *Research Policy*, 27(9):901–913, 1998.

[275] P. Langley and S. Sage. Induction of selective Bayesian classifiers. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 399–406, 1994.

[276] P. Larrañaga, C. M. H. Kuijpers, R. H. Murga, and Y. Yurramendi. Learning Bayesian network structures by searching for the best ordering with genetic algorithms. *IEEE Transactions on System, Man and Cybernetics. Part A: Systems and Humans*, 26(4):487–493, 1996.

[277] P. Larrañaga, M. Poza, Y. Yurramendi, R. H. Murga, and C. M. H. Kuijpers. Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):912–926, 1996.

[278] S. C. Larson. The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22(1):45–55, 1931.

[279] E. M. Lasda Bergman. Finding citations to social work literature: The relative benefits of using Web of Science, Scopus, or Google Scholar. *The Journal of Academic Librarianship*, 38(6):370–379, 2012.

[280] S. L. Lauritzen. *Graphical Models*. Clarendon Press, 1996.

[281] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, 50(2):157–224, 1988.

[282] S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17(1):31–57, 1989.

[283] S. Lehmann, A. D. Jackson, and B. E. Lautrup. Measures for measures. *Nature*, 444(7122):1003–1004, 2006.

[284] M. Levine-Clark and E. L. Gil. A comparative citation analysis of Web of Science, Scopus, and Google Scholar. *Journal of Business and Finance Librarianship*, 14(1):32–46, 2009.

[285] J. M. Levitt and M. Thelwall. A combined bibliometric indicator to predict article impact. *Information Processing and Management*, 47(2):300–308, 2011.

[286] J.M. Levitt and M. Thelwall. Citation levels and collaboration within library and information science. *Journal of the American Society for Information Science and Technology*, 60(3):434–442, 2009.

[287] L. Leydesdorff. How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology*, 60(7):1327–1336, 2009.

[288] L. Leydesdorff, F. de Moya-Anegón, and V. P. Guerrero-Bote. Journal maps on the basis of Scopus data: A comparison with the journal citation reports of the ISI. *Journal of the American Society for Information Science and Technology*, 61(2):352–369, 2010.

[289] L. Leydesdorff and S. Milojevic. Scientometrics. *Forthcoming in: International Encyclopedia of Social and Behavioral Sciences*, 2015.

[290] L. Liang. H-index sequence and h-index matrix: Constructions and applications. *Scientometrics*, 69(1):153–159, 2006.

[291] C. H. Liao. How to improve research quality? Examining the impacts of collaboration intensity and member diversity in collaboration networks. *Scientometrics*, 86(3):747–761, 2011.

[292] H. T. Lin and L. Li. Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 24(5):1329–1367, 2012.

[293] D. Lindsey. Production and citation measures in the sociology of science: The problem of multiple authorship. *Social Studies of Science*, 10(2):145–162, 1980.

[294] C. X. Ling, Q. Yang, J. Wang, and S. Zhang. Decision trees with minimal costs. In *Proceedings of the 21st International Conference on Machine Learning*, pages 69–77, 2004.

[295] G. Liu. *Introduction to Combinatorial Mathematics*. McGraw-Hill, 1968.

[296] C. Lokker, K. A. McKibbon, and R. J. McKinlay. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: Retrospective cohort study. *British Medical Journal*, 336:655–657, 2008.

[297] S. Lomax and S. Vadera. A survey of cost-sensitive decision tree induction algorithms. *ACM Computing Surveys*, 45(2):Article 16, 2013.

[298] S. López-Berna, N. Papí-Gálvez, and M. Martín-Llaguno. Productividad científica en España sobre las profesiones de comunicación entre 1971 y 2009. *Revista Española de Documentación Científica*, 34(2):212–231, 2011.

[299] R. López de Mántaras and E. Armengol. Machine learning from examples: Inductive and lazy methods. *Data and Knowledge Engineering*, 25(1-2):99–123, 1998.

[300] J. Lundberg. Lifting the crown-citation z-score. *Journal of Informetrics*, 1(2):145–154, 2007.

[301] N. Ma, J. Guan, and Y. Zhao. Bringing PageRank to the citation analysis. *Information Processing and Management*, 44(2):800–810, 2008.

[302] Y. S. Maarek and I. Z. Ben Shaul. Automatically organizing bookmarks per contents. *Computer Networks and ISDN Systems*, 28(7-11):1321–1333, 1996.

[303] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[304] M. H. MacRoberts and B. R. MacRoberts. Problems of citation analysis. *Scientometrics*, 36(3):435–444, 1996.

[305] M. H. Magri and A. Solari. The SCI Journal Citation Reports: A potential tool for studying journals? I. Description of the JCR journal population based on the number of citations received, number of source items, impact factor, immediacy index and cited half-life. *Scientometrics*, 35(1):93–117, 1996.

[306] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, pages 49–55, 1936.

[307] M. Manafy. Scopus of influence: Content selection committee announced. *Econtent*, 28(10):12, 2005.

[308] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[309] T. Marchant. Score-based bibliometric rankings of authors. *Journal of the American Society for Information Science and Technology*, 60(6):1132–1137, 2009.

[310] B. Martin. The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36(3):343–362, 1996.

[311] M. Martínez-Morales, R. Garza-Domínguez, N. Cruz-Ramírez, A. Guerra-Hernández, and J. L. Jiménez-Andrade. A method based on genetic algorithms and fuzzy logic to induce Bayesian networks. In *Proceedings of the Fifth Mexican International Conference in Computer Science*, pages 176–180, 2004.

[312] C. McCarty, J. W. Jawitz, A. Hopkins, and A. Goldman. Predicting author h-index using characteristics of the co-author network. *Scientometrics*, 96(2):467–483, 2013.

[313] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B*, 42(2):109–142, 1980.

[314] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1983.

[315] W. McCulloch and W. Pitts. A logical calculus of ideas imminent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[316] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

[317] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[318] G. J. McLachlan, D. Peel, K. Basford, and P. Adams. Fitting of mixtures of normal and t-components. *Journal of Statistical Software*, 4(2):909–927, 1999.

[319] M. E. McVeigh and S. J. Mann. The journal impact factor denominator defining citable (counted) items. *Journal of the American Medical Association*, 302(10):1107–1109, 2009.

[320] L. I. Meho and C. R. Sugimoto. Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, 59(11):1711–1726, 2008.

[321] L. I. Meho and K. Yang. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13):2105–2125, 2007.

[322] S. Menard. *Applied Logistic Regression Analysis*. Sage, 2002.

[323] S. Mikki. Comparing Google Scholar and ISI Web of Science for earth sciences. *Scientometrics*, 82(2):321–331, 2010.

[324] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[325] M. Minsky. Steps toward artificial intelligence. *IRE*, 49(1):8–30, 1961.

[326] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

[327] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[328] H. F. Moed. *Citation Analysis in Research Evaluation*. Springer, 2005.

[329] H. F. Moed. Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4(3):265–277, 2010.

[330] H. F. Moed, R. E. De Bruin, and T. N. van Leeuwen. New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics*, 33(3):381–422, 1995.

[331] H. F. Moed, T. N. van Leeuwen, and J. Reedijk. Towards appropriate indicators of journal impact. *Scientometrics*, 46(3):575–589, 1999.

[332] A. Molinari and J. Molinari. Mathematical aspects of a new criterion for ranking scientific institutions based on the h-index. *Scientometrics*, 75(2):339–356, 2008.

[333] F. Mosteller and J. W. Tukey. *Data Analysis, Including Statistics.* Addison-Wesley, 1968.

[334] H. Moxham and J. Anderson. Peer review. A view from the inside. *Science and Technology Policy*, 5(1):7–15, 1992.

[335] F. Moya-Anegón, Z. Chinchilla-Rodríguez, E. Corera-Álvarez, M. Gómez-Crisóstomo, A. González-Molina, and F.J. Muñoz-Fernández. La productividad ISI de las universidades españolas (2000-2004). *El profesional de la información*, 16(4):354–358, 2007.

[336] F. Moya-Anegón, Z. Chinchilla-Rodríguez, E. Corera-Álvarez, B. Vargas-Quesada, F. Muñoz-Fernández, and V. Herrero-Solana. Análisis de dominio institucional: La producción científica de la Universidad de Granada (SCI 1991-99). *Revista Española de Documentación Científica*, 28(2):170–195, 2005.

[337] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

[338] A. Mulligan. Is peer review in crisis? *Oral Oncology*, 41:135–141, 2005.

[339] S. K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4):345–389, 1998.

[340] J. W. Myers, K. B. Laskey, and K. A. DeJong. Learning Bayesian networks from incomplete data using evolutionary algorithms. In *15th Conference on Uncertainty in Artificial Intelligence*, pages 476–485, 1999.

[341] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown. An introduction to decision tree modeling. *Journal of Chemometrics*, 18(6):275–285, 2004.

[342] V. V. Nalimov and Z. M. Mulchenko. *Naoukometriia. Izuchenie Razvitiia Nauki kak Informatsionvo Prosessa (Scientometrics. Study of the Development of Science as an Information Process).* Nauka, 1969.

[343] F. Narin. *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity.* Computer Horizons, 1976.

[344] R. E. Neapolitan. *Learning Bayesian Networks.* Prentice Hall, 2003.

[345] C. Neocleous and C. Schizas. Artificial neural network learning: A comparative review. In *Lecture Notes in Artificial Intelligence*, volume 2308, pages 300–313. Springer-Verlag, 2002.

[346] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. McGraw-Hill, 1996.

[347] M. Norris and C. Oppenheim. Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, 1(2):161–169, 2007.

[348] C. Olmeda-Gómez, M.A. Ovalle-Parandones, A. Perianes-Rodríguez, and F. Moya-Anegón. Impacto internacional de la investigación y la colaboración científica de las Universidades de Cataluña. 2000-2004. *Revista Española de Documentación Científica*, 31(4):591–611, 2008.

[349] G. M. Olson and J. S. Olson. Distance matters. *Human Computer Interaction*, 15(2):139–179, 2000.

[350] T. Opthof and L. Leydesdorff. Caveats for the journal and field normalizations in the CWTS ("leiden") evaluations of research performance. *Journal of Informetrics*, 4(3):423–430, 2010.

[351] E. Pain. Research cuts will cause "exodus" from Spain. *Science*, 336(6078):139–140, 2012.

[352] D. Palomares-Montero and A. García-Aracil. Fuzzy cluster analysis on Spanish public universities. In *Investigaciones de Economía de la Educación*, volume 5, chapter 49, pages 976–994. Asociación de Economía de la Educación, 2010.

[353] J. Panaretos and M. Chrisovaladis. Assessing scientific research performance and impact with single indices. *Scientometrics*, 81(3):635–670, 2008.

[354] M. J. Pazzani. Searching for dependencies in Bayesian classifiers. In *Learning from data: Artificial Intelligence and Statistics V*, volume 112, pages 239–248. Springer, 1996.

[355] J. Peña, J. Lozano, and P. Larrañaga. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*, 20(10):1027–1040, 1999.

[356] J. Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society*, pages 329–334, 1985.

[357] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, 1988.

[358] J. Pearl. A theory of inferred causation. In *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning*, pages 441–452, 1991.

[359] K. Pearson. On lines and planes of closest fit to systems of point in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[360] K. Pearson. Notes on the history of correlation. *Biometrika*, 13(1):25–45, 1920.

[361] T. Perneger. Relation between online "hit counts" and subsequent citations: Prospective study of research papers in the BMJ. *British Medical Journal*, 329(7465):546–547, 2004.

[362] O. Persson, W. Glanzel, and R. Danell. Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, 60(3):421–432, 2004.

[363] G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing and Management*, 12(5):297–312, 1976.

[364] B.L. Ponomariov and P.C. Boardman. Influencing scientists' collaboration and productivity patterns through new institutions: University research centers and scientific and technical human capital. *Research Policy*, 39(5):613–624, 2010.

[365] R. Potharst and J. C. Bioch. Decision trees for ordinal classification. *Intelligent Data Analysis*, 4(2):97–112, 2000.

[366] O. Pourret, P. Naim, and B. Marcot. *Bayesian Networks: A Practical Guide to Applications*. Wiley, 2008.

[367] G. Prathap. Hirsch-type indices for ranking institutions' scientific research output. *Current Science*, 91(11):1439, 2006.

[368] G. Prathap. The iCE approach for journal evaluation. *Scientometrics*, 85(2):561–565, 2010.

[369] G. Prathap. The fractional and harmonic p-indices for multiple authorship. *Scientometrics*, 86(2):239–244, 2011.

[370] S. Presser. Collaboration and the quality of research. *Social Studies of Science*, 10(1):95–101, 1980.

[371] A. Pritchard. Statistical bibliography or bibliometrics? *Journal of Documentation*, 25(4):348–349, 1969.

[372] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, USA, 1993.

[373] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Towards an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.

[374] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[375] C. R. Rao and H. Toutenburg. *Linear Models: Least Squares and Alternatives*. Springer, 1995.

[376] J. Rao and A. Shao. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79(4):811–822, 1992.

[377] M. J. Reyes-Barragán, V. P. Guerrero-Bote, and F. Moya-Anegón. Proyección internacional de la investigación de Extremadura (1990-2002). *Revista Española de Documentación Científica*, 29(4):525–550, 2006.

[378] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 2008.

[379] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[380] A. Rodríguez-Navarro. A simple index for the high-citation tail of citation distribution to quantify research performance in countries and institutions. *PLoS ONE*, 6(5):e20510, 2011.

[381] R. Rojo and I. Gómez. Analysis of the Spanish scientific and technological output in the ICT sector. *Scientometrics*, 66(1):101–121, 2006.

[382] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[383] F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, 1962.

[384] R. Rousseau. New developments related to the Hirsch index. *Science Focus*, 1(4):23–25, 2006.

[385] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[386] F. Ruane and R. S. J. Tol. Rational (successive) h-indices: An application to economics in the Republic of Ireland. *Scientometrics*, 75(2):395–405, 2008.

[387] D. Rumerlhart, G. Hinton, and R. Williams. Learning internal representations by backpropagation errors. *Nature*, 323:533–536, 1986.

[388] P. Russel and T. Rao. On habitat and association of species of Anopheline Larvae in South-Eastern Madras. *Journal of Malaria Institute India*, 3(1):153–178, 1940.

[389] G. Saad. Convergent validity between metrics of journal prestige: The eigenfactor, article influence, h-index scores, and impact factors. *Journal of the American Society for Information Science and Technology*, In press, 2015.

[390] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[391] M. Sahami. Learning limited dependence Bayesian classifiers. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 335–338, 1996.

[392] M. Sanderson. Revisiting h measured on UK LIS academics. *Journal of the American Society for Information Science and Technology*, 59(7):1184–1190, 2008.

[393] T. Scarpa. Peer review at NIH. *Science*, 311(5757):41, 2006.

[394] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[395] M. Schreiber. Self-citation corrections for the Hirsch index. *Euro Physics Letters*, 78(3):30002, 2007.

[396] M. Schreiber. An empirical investigation of the g-index for 26 physicists in comparison with the h-index, the a-index, and the r-index. *Journal of the American Society for Information Science and Technology*, 59(9):1513–1522, 2008.

[397] M. Schreiber. A modification of the h-index: The hm-index accounts for multi-authored manuscripts. *Journal of Informetrics*, 2(3):211–216, 2008.

[398] M. Schreiber. To share the fame in a fair way, hm for multi-authored manuscripts. *New Journal of Physics*, 10(040201):1–9, 2008.

[399] M. Schreiber. Fractionalized counting of publications for the g-index. *Journal of the American Society for Information Science and Technology*, 60(10):2145–2150, 2009.

[400] A. Schubert. Successive h-indices. *Scientometrics*, 70(1):201–205, 2007.

[401] A. Schubert. Using the h-index for assessing single publications. *Scientometrics*, 78(3):559–565, 2009.

[402] A. Schubert and T. Braun. Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, 9(5-6):281–291, 1986.

[403] A. Schubert and W. Glanzel. A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, 1(3):179–184, 2007.

[404] A. Schubert, W. Glanzel, and T. Braun. Relative citation rate: A new indicator for measuring the impact of publications. In *Proceedings of the First National Conference with International Participation on Scientometrics and Linguistics of Scientific Text*, pages 80–81, 1983.

[405] G. Schwarz. Estimating the dimensions of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[406] G. A. F. Seber and A. J. Lee. *Linear Regression Analysis*. John Wiley and Sons, 2012.

[407] P. O. Seglen. Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314(7079):498–502, 1997.

[408] M. Sember, A. Utrobicic, and J. Petrak. Croatian medical journal citation score in Web of Science, Scopus, and Google Scholar. *Croatian Medical Journal*, 51(2):99–103, 2010.

[409] A. Serenko. The development of an AI journal ranking based on the revealed preference approach. *Journal of Informetrics*, 4(4):447–459, 2010.

[410] R. D. Shachter. Intelligent probabilistic inference. In *Proceedings of the First Annual Conference on Uncertainty in Artificial Intelligence*, pages 371–382, 1985.

[411] R. D. Shachter and C. R. Kenley. Gaussian influence diagrams. *Management Science*, 35(5):527–550, 1989.

[412] A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. In *Advances in Neural Information Processing Systems*, volume 15, pages 961–968. MIT Press, 2003.

[413] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[414] V. S. Sheng and C. X. Ling. Roulette sampling for cost-sensitive learning. In *Proceedings of the 18th European conference on Machine Learning*, Lecture Notes in Computer Science, pages 724–731. Springer, 2007.

[415] A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos. Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2):253–280, 2007.

[416] R. Sivaraj and T. Ravichandran. A review of selection methods in genetic algorithms. *International Journal of Engineering Science and Technology*, 3(5):3792–3797, 2011.

[417] C. A. B. Smith. Some examples of discrimination. *Annals of Eugenics*, 13(1):272–282, 1946.

[418] P. W. F. Smith and J. Whittaker. Edge exclusion tests for graphical Gaussian models. In *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, pages 555–574, 1998.

[419] L. Smolinsky and A. Lercher. Citation rates in mathematics: A study of variation by subdiscipline. *Scientometrics*, 91(3):911–924, 2012.

[420] P. Sneath. The applications of computers to taxonomy. *Journal of General Microbiology*, 17(1):201–206, 1957.

[421] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28(22):1409–1438, 1958.

[422] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence*, pages 1015–1021, 2006.

[423] J. M. Soler. A rational indicator of scientific creativity. *Journal of Informetrics*, 1(2):123–130, 2007.

[424] J. Solomon. Programmers, professors, and parasites: Credit and co-authorship in computer science. *Science and Engineering Ethics*, 14(5):476–489, 2009.

[425] R. Sooryamoorthy. Do types of collaboration change citation? Collaboration and citation patterns of South African science publications. *Scientometrics*, 81(1):177–193, 2009.

[426] T. Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyzes of the vegetation on Danish commons. *Biologiske Skrifter*, 5(1):1–34, 1948.

[427] K. A. Spackman. Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 160–163, 1989.

[428] W. M. Spears and V. Anand. A study of crossover operators in genetic programming. In *Proceedings of the 6th International Symposium on Methodologies for Intelligent Systems*, pages 409–418, 1991.

[429] D. J. Spiegelhalter. Probabilistic reasoning in predictive expert systems. In *Proceedings of the First Annual Conference on Uncertainty in Artificial Intelligence*, pages 47–68, 1985.

[430] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

[431] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Springer, 1993.

[432] S. V. Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1):77–89, 1997.

[433] M. Stone. Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistic Society*, 36:111–147, 1974.

[434] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[435] J. M. Tague-Sutcliffe. An introduction to informetrics. *Information Processing and Management*, 28(1):1–3, 1992.

[436] K. M. Ting. Inducing cost-sensitive trees via instances weighting. In *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 23–26, 1998.

[437] D. Torres-Salinas, E. Delgado López-Cózar, and E. Jiménez-Contreras. Análisis de la producción de la Universidad de Navarra en revistas de Ciencias Sociales y Humanidades empleando rankings de revistas españolas y la Web of Science. *Revista Española de Documentación Científica*, 32(1):22–39, 2009.

[438] D. Torres-Salinas, E. Lopez-Cózar, and E. Jiménez-Contreras. Ranking of departments and researchers within a university using two different databases: Web of Science versus Scopus. *Scientometrics*, 80(3):761–774, 2009.

[439] D. Torres-Salinas, J. G. Moreno-Torres, E. Delgado-López-Cózar, and F. Herrera. A methodology for institution-field ranking based on a bidimensional analysis: the IFQ^2A index. *Scientometrics*, 88(3):771–786, 2011.

[440] D. Torres-Salinas, J. G. Moreno-Torres, N. Robinson-García, E. Delgado-López-Cózar, and F. Herrera. Rankings ISI of Spanish universities according to fields and scientific disciplines (2nd ed. 2011). *El Profesional de la Información*, 20(6):701–709, 2011. In Spanish.

[441] A. Tucker, X. Liu, and A. Ogden-Swift. Evolutionary learning of dynamic probabilistic models with large time lags. *International Journal of Intelligent Systems*, 16(5):621–645, 2001.

[442] P. D. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 2:369–409, 1995.

[443] C. Urbano, A. Borrego, J.M. Brucart, A. Cosculluela, and M. Somoza. Análisis bibliométrico de la bibliografía citada en estudios de filología española. *Revista Española de Documentación Científica*, 28(4):439–461, 2005.

[444] M. Vallejo Ruiz, A. Fernández Cano, and M. Torralbo Rodríguez. Patrones de citación en la investigación española en educación matemática. *Revista Española de Documentación Científica*, 29(3):382–397, 2006.

[445] N. J. van Eck and L. Waltman. Generalizing the h- and g-indices. *Journal of Informetrics*, 2(4):263–271, 2008.

[446] R. van Engelen. Approximating Bayesian belief networks by arc removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(8):916–920, 1997.

[447] G. van Hooydonk. Fractional counting of multi-authored publications: consequences for the impact of authors. *Journal of the American Society for Information Science*, 48(10):944–945, 1997.

[448] T. van Leeuwen. Testing the validity of the Hirsch-index for research assessment purposes. *Research Evaluation*, 17(2):157–160, 2008.

[449] T. N. van Leeuwen, H. F. Moed, and J. Reedijk. Critical comments on institute for scientific information impact factors: A sample of inorganic molecular chemistry journals. *Journal of Information Science*, 25(6):189–198, 1999.

[450] A. F. J. van Raan. Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, 36(3):397–420, 1996.

[451] A. F. J. van Raan. The influence of international collaboration on the impact of research result. *Scientometrics*, 42(3):423–428, 1998.

[452] A. F. J. van Raan. Special topic issue: Science and technology indicators. *Journal of the American Society for Information Science*, 49(1):3–81, 1998.

[453] A. F. J. van Raan. Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1):133–143, 2005.

[454] A. F. J. van Raan. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3):491–502, 2006.

[455] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

[456] J. Vanclay. On the robustness of the h-index. *Journal of the American Society for Information Science and Technology*, 58(10):1547–1550, 2007.

[457] J. K. Vanclay. Bias in the journal impact factor. *Scientometrics*, 78(1):3–12, 2009.

[458] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.

[459] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[460] D. Vidaurre, C. Bielza, and P. Larrañaga. Learning an l1-regularized Gaussian Bayesian network in the equivalence class space. *IEEE Transactions os Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(5):1231–1242, 2010.

[461] E. S. Vieira, J. A. S. Cabral, and J. A. N. F. Gomez. Definition of a model based on bibliometric indicators for assessing applicants to academic positions. *Journal of the Association for Information Science and Technology*, 65(3):560–577, 2014.

[462] E. S. Vieira, J. A. S. Cabral, and J. A. N. F. Gomez. How good is a model based on bibliometric indicators in predicting the final decisions made by peers? *Journal of Informetrics*, 8(2):390–405, 2014.

[463] P. Vinkler. Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics*, 10(3-4):157–177, 1986.

[464] P. Vinkler. Eminence of scientists in the light of the h-index and other scientometric indicators. *Journal of Information Science*, 33(4):481–491, 2007.

[465] P. Vinkler. Indicators are the essence of scientometrics and bibliometrics. *Scientometrics*, 85(3):861–866, 2010.

[466] J. Wainer, E. C. Xavier, and F. Bezerra. Scientific production in computer science: A comparative study of Brazil and other countries. *Scientometrics*, 81(2):535–547, 2009.

[467] C. Wallace and D. Dowe. Intrinsic classification by MML - The SNOB program. In *Proceeding of the 7th Australian Joint Conference on artificial intelligence*, pages 37–44, 1994.

[468] L. Waltman and N. J. van Eck. The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, 63(2):406–415, 2012.

[469] L. Waltman, N. J. van Eck, T. N. van Leeuwen, M. S. Visser, and A. F. J. van Raan. Towards a new crown indicator: An empirical analysis. *Scientometrics*, 87(3):467–481, 2011.

[470] J. K. Wan, P. H. Hua, and R. Rousseau. The pure h-index: Calculating an author's h-index by taking co-authors into account. *Collnet Journal of Scientometrics and Information Management*, 1(2):1–5, 2007.

[471] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*, 58(301):236–244, 1963.

[472] A. Watson. Comparing citations and downloads for individual articles. *Journal of Vision*, 9(4):1–4, 2009.

[473] P. Weingart. Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(2):117–131, 2005.

[474] S. Weisberg. *Applied Linear Regression*. John Wiley and Sons, 2014.

[475] M. P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44(3):257–303, 1990.

[476] M. C. Wendl. H-index: However ranked, citations need context. *Nature*, 449(7161):403, 2007.

[477] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, 1990.

[478] I. H. Witten and E. Frank. *Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, 2005.

[479] G. J. Woeginger. An axiomatic characterization of the Hirsch-index. *Mathematical Social Sciences*, 56(2):224–232, 2008.

[480] Q. Wu. The w-index: A measure to assess scientific impact by focusing on widely cited papers. *Journal of the American Society for Information Science and Technology*, 61(3):609–614, 2010.

[481] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[482] F. Y. Ye and R. Rousseau. The power law model and total career h-index sequences. *Journal of Informetrics*, 2(4):288–297, 2008.

[483] R. Yehezkel and B. Lerner. Bayesian network structure learning by recursive autonomy identification. *Journal of Machine Learning Research*, 10:1527–1570, 2009.

[484] K. Yeung, D. Hayunor, and W. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, 2001.

[485] G. Yu and L. Wang. The self-cited rate of scientific journals and the manipulation of their impact factors. *Scientometrics*, 73(3):320–330, 2007.

[486] B. Zadrozny and C. Elkan. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, pages 204–213, 2001.

[487] B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate instance weighting. In *Proceedings of the 3rd International Conference on Data Mining*, pages 435–442, 2003.

[488] R. Zetterstrom. The number of authors of scientific publications. *Acta Paediatrica*, 93(5):581–582, 2004.

[489] C. T. Zhang. The e-index, complementing the h-index for excess citations. *PLoS ONE*, 4(5):e5429, 2009.

[490] X. Zhu and A. Goldberg. *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers, 2009.

[491] M. Zitt and E. Bassecoulard. Challenges for scientometric indicators: data demining, knowledge-flow measurements and diversity issues. *Ethics in Scice and Environmental Politics*, 8(1):49–60, 2008.

[492] M. Zitt and H. Small. Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology*, 59(11):1856–1860, 2008.

[493] H. Zuckerman. Patterns of name-ordering among authors of scientific papers: a study of social symbolism and its ambiguity. *American Journal of Sociology*, 74(3):276–291, 1968.