Predicting the h-index with cost-sensitive naive Bayes

Alfonso Ibáñez, Pedro Larrañaga and Concha Bielza
Computational Intelligence Group, Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid, Madrid, Spain
{*aibanez,plarranaga,mcbielza*}*@fi.upm.es*

***Abstract*—Bibliometric indices are an increasingly important topic for the scientific community nowadays. One of the most successful bibliometric indices is the well-known *h-index*. In view of the attention attracted by this index, our research is based on the construction of several prediction models to forecast the *h-index* of Spanish professors (with a permanent position) for a four-year time horizon. We built two different types of models (junior models and senior models) to differentiate between professors' seniority. 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. Results 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.**

***Keywords*-cost-sensitive naive Bayes; h-index; bibliometric indices;**

I. Introduction

Bibliometric indices are quantitative metrics for evaluating and comparing the research activity of individual researchers according to their output. Bibliometric indices are an increasingly important topic for the scientific community nowadays. In fact, many funding agencies and promotion committees use them for accepting research projects and contracting researchers, among others.

Several bibliometric indices have been developed (see reviews [1], [2]). One of the most successful indices was proposed by Jorge Hirsch and it was called the *h-index* [3]. 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. The *h-index* is based on a list of publications ranked in descending order by the times cited. The value of *h* is equal to the number of papers (N) in the list that have N or more citations. 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 741 times (May, 2011).

Some bibliometric indices have been predicted in the literature (e.g., [4], [5]). These predictions used time series modeled by exponential and exponential smoothing functions. Other methods, like Bayesian networks, logistic regression, decision trees and the K-NN algorithm were also used for making predictions [6]. The above papers and our work have a similar aim, but despite this, the class variable, the predictive features and the used methods are different.

Focusing on the *h-index*, we noted that there are not many papers related to the prediction of this bibliometric index. The power law model [7] was used to analyze the *h-index* as a function of time [8]. Nonlinear regression was also used to predict the *h-index* of authors, journals and universities [9]. 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.

The interest and originality of our study 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, we build prediction models to forecast the annual increase of the *h-index* of Spanish professors. All the professors belong to public universitites and are associated with three specific areas: Computer Architecture and Technology, Computer Science and Artificial Intelligence, and Computer Languages and Systems. Finally, we considered the *h-index* prediction as an ordinal classification problem. For this reason, we are concerned not only with maximizing the classification accuracy, but also with minimizing the weighted distances between the actual and the predicted values.

The remainder of the paper is organized as follows. Section 2 explains some methods (classification methods, feature selection and assessment procedure), and reviews some basic concepts about bibliometric indices on which our work is based. Section 3 presents the dataset used, and the different models learned. Finally, Section 4 contains some conclusions emphasizing the original contribution of the paper and future research on the topic.

II. Methods

A. Supervised classification method

The cost-sensitive naive Bayes is an adaptation of the original naive Bayes [10] 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 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 *val(C)*. Figure 1 represents the naive

978-1-4577-1676-8/11/$26.00

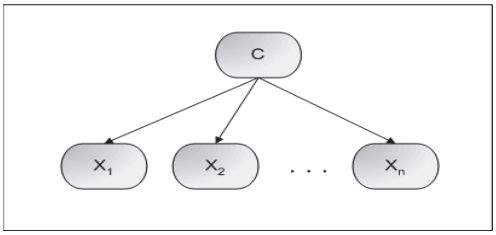

Figure 1. Naive Bayes structure

Bayes structure. 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 **x**:

$$c^* = \arg \min_{c \in val(C)} \sum_{c'=1}^{val(C)} p(c' \mid \mathbf{x}) \;\; cost(c \mid c')$$

where

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

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

B. Predictive features

In this section, we describe the dataset structure. The dataset structure of each model is made up of *area*, *position*, *university*, *seniority* and 12 *bibliometric indices*. In the following, we explain each feature.

X_1) *Area:* This feature represents the 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).

X_2) *Position:* This feature corresponds to the position of each professor. It has possible four values (Full professor, Associate professor (type I), Associate professor (type II) and Assistant professor).

X_3) *University:* This feature 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,...).

X_4) *Seniority:* This feature 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.

X_5) *Documents:* This is an index associated with the number of papers published by each professor. This index represents the productivity of each specific professor.

X_6) *Citations:* This is an index associated with the number of citations received by each professor. This index represents the visibility of each specific professor.

X_7) *The h-index:* The *h-index* is proposed in [3].

X_8) *The 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*, proposed in [11], 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. 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.

X_9) *The hg-index:* A new index called the *hg-index*, which is based on the *h-index* and the *g-index*, is presented in [12]. It intends to provide a more balanced view of the scientific production of researchers. The *hg-index* of a researcher is computed as the geometric mean of its *h-index* and *g-index*, that is,

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

where h corresponds to the value of the *h-index*, and g corresponds to the value of the *g-index*.

X_{10}) *The a-index:* The *a-index* was proposed in [13]. This index is calculated for papers that are in the *h-core* only, that is, the first h papers. It is defined as the average number of citations received by the articles included in the *h-core*. This index measures the citation intensity in the *h-core*. The *a-index* can be very sensitive to just a very few papers receiving extremely high citation counts.

X_{11}) *The 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*, is proposed in [14] as a variation on the *a-index*. This index, which was designed to illustrate the impact of the papers in the *h-core*, is the median number of citations received by papers in the *h-core*.

X_{12}) *The q^2-index:* A new index, called q^2-*index*, is developed in [15] to provide a more global view of scientific production. This index 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), that is,

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

where h corresponds to the value of the *h-index*, and m corresponds to the value of the *m-index*.

X_{13}) *The h_r-index:* The *rational h-index*, which is an extension of the original *h-index*, was proposed in [16]. This index takes 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

$$rational\ h\text{-}index = (h+1) - \frac{Cit(h+1)}{2h+1},$$

where h is the value of the *h-index*, and *Cit(h+1)* is the number of citations received by article *h+1*.

X_{14}) *The h_i-index:* The *individual h-index*, proposed in [17], is complementary to the *h-index* and indicates 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 for this procedure is to measure the effective individual average productivity.

$$individual\ h\text{-}index = \frac{h}{N_a},$$

where h is the value of the *h-index*, and N_a is the mean number of authors in the h papers.

X_{15}) *The c-index:* This index was proposed in [18]. It is based on creativity, defined as the creation of new scientific knowledge. Its objective is to highlight papers that receives many citations and have few references. This index is easily calculated from the citations and references of the author's papers:

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

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 published 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; $z = (m-1)/(n+5)$; and A, B, a, b are arbitrary parameters.

X_{16}) *The h_c-index:* The original *h-index* cannot distinguish between inactive scientists, young scientists and senior scientists, who are still contributing nowadays. For this reason, there is a need to define a new index that takes into account the "age" of papers. A novel score *Sc(i)* is defined for a paper *i* based on citation counting:

$$Sc(i) = \gamma \cdot (Y(now) - Y(i) + 1)^{-\delta} Cit(i),$$

where *Y(now)* is the current year, *Y(i)* is the publication year of paper *i*; *Cit(i)* is the number of citations received by paper *i*; γ and δ are arbitrary parameters.

Using the above score, the value of old papers gradually declines, even if they still receive citations. Therefore, a new *contemporary h-index* is defined in [19]. Its definition states that "a researcher has index h_c, if h_c of its published papers gets a score of *Sc(i)* $\geq h_c$ each, and the other papers get a score of *Sc(i)* $< h_c$".

C. Selecting 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 [20]. In this case, we used correlation-based feature selection (CFS) as our 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.

D. Assessment procedure

We chose *k*-fold cross-validation as the procedure for estimating the accuracy of models classifying new cases according to the value of the predictive features. This method divides all cases from the dataset into *k* 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* percentages of well-classified cases are averaged to output the estimated value of the model learned from all cases to classify new cases [21]. In this paper, we used 10 times 10-fold cross-validation as the assessment procedure. In this way, it is possible to perform a statistical hypothesis test for indicating if some values are statistically better than others at a specified significance level.

Table I
DATA DISTRIBUTION OF JUNIOR MODELS

	First-year	Second-year	Third-year	Fourth-year
Δh=0	239	205	159	146
Δh=1	50	82	119	125
Δh=2	-	2	10	17
Δh=3	-	-	1	1

III. RESULTS

A. Dataset construction

Our work is based on the construction of predictive models to forecast the *h-index* of Spanish professors for a four-year time horizon. For this study we focus on papers published by each professor from January 1, 1978 to December 31, 2005. Using this information, we built two different types of predictive models: senior models and junior models. 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 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 taken into account. In the following, we illustrate the different phases for building the predictive models.

The first step was to submit a request to the Spanish Ministry of Education for a list of professors associated with three specific areas (Computer Architecture and Technology, Computer Science and Artificial Intelligence, and finally, Computer Languages and Systems) until December 31, 2009. This list includes the full name of each professor (2004 professors), and their associated university, position and research area. The next step was to obtain the publication list and citation data (until December 31, 2009) for each professor. This information was downloaded from the Web of Science (ISI Web of Knowledge). The last step was to use all this information to calculate some bibliometric indices associated with the selected professors. These bibliometric indices are: *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 *c-index* and the h_c*-index*.

Table II
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*	√	√		

B. Models

1) Data distribution: After collecting the publication list and citation data of all professors, we observe that junior models have been learnt from data on 289 professors, whereas senior models have been learnt from data on 352 professors.

Table I shows the distribution of the professors selected in junior models according to their annual increase of the *h-index* value within the first four years. Taking the first year as an example, we observe that most professors (239 cases) have Δ*h-index*=0, whereas only 50 professors have Δ*h-index*=1. We also show that most junior professors have Δ*h-index*=0 in the second, third and fourth year. On the other hand, 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. The distribution associated with senior professors is not shown for space reasons.

2) Feature selection: In order to determine if all the predictive features (*area*, *position*, *university*, *seniority* and 12 *bibliometric indices*) are equally important or necessary for discriminating between the different values of the annual increase of the *h-index*, we performed feature selection.

Table II shows the predictive features that are selected in senior models after running CFS for feature 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.

We learnt two cost-sensitive naive Bayes models with the intention of checking the benefits of applying feature selection (e.g. better classification performance). Each model is learnt from a different dataset. The first dataset (Dataset$_{nofs}$) does not include a feature selection whereas the second

Table III
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)	68.85 ± 5.78 - - (16)
Dataset $_{fs}$	81.31 ± 2.37 - - (1)	69.52 ± 5.58 - - (9)
Second-year	3 classes	7 classes
Dataset $_{nofs}$	47.74 ± 13.96 - - (16)	57.15 ± 6.68 - - (16)
Dataset $_{fs}$	71.46 ± 5.72 * - - (2)	58.20 ± 6.78 - - (8)
Third-year	4 classes	12 classes
Dataset $_{nofs}$	54.08 ± 8.23 - - (16)	53.58 ± 7.81 - - (16)
Dataset $_{fs}$	55.02 ± 6.94 - - (2)	51.02 ± 6.49 - - (5)
Fourth-year	4 classes	15 classes
Dataset $_{nofs}$	48.92 ± 7.23 - - (16)	47.85 ± 8.19 - - (16)
Dataset $_{fs}$	50.21 ± 7.90 - - (3)	48.51 ± 7.35 - - (5)

dataset (Dataset$_{fs}$) does.

The 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. We used other cost matrices (quadratic functions) but the results (not shown) did not vary too much.

Table III lists the accuracy and the standard deviations of predictive models. Also, it 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, we show 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 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. Table III shows that most of the models obtain better classification performance when feature selection is performed. Finally, we note 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.

The symbol (*) placed beside a result indicates that it is statistically (*t-test*) better than the other model at a specified significance level of 0.05.

3) Accuracy and average cost: In order to determine if the accuracy values are reasonable, we compare cost-sensitive naive Bayes with the standard formulation of naive Bayes. Table IV 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, we show 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, we note that 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.

We enlarged our datasets by also including the bibliometric indices for the two and three previous years to again build junior and senior models. The estimation of the accuracy and cost obtained using these datasets (results are not shown) were very similar to the values shown in Table IV.

Other models are built for predicting the exact value of the *h-index* instead of the *h-index* increase. These models are also learnt from the same predictive features and classification method. The results of these models (not shown for space reasons) were similar to the values shown in Table IV.

4) Example: We predict the increase of the *h-index* value of a new senior professor in the first year. Table V shows the parameters that define the model. 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 ($p(x_i|c)$). We use the Laplace estimator to compute the conditional distributions of discrete features. Table V does not show all the parameters for space reasons.

Given a senior professor (**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 V. 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

Table IV
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)	69.50 ± 5.59 - - (0.412)
NB$_{cost}$	81.31 ± 2.37 - - (0.187)	69.52 ± 5.58 - - (0.398)
Second-year	3 classes	7 classes
NB	71.29 ± 5.68 - - (0.294)	58.20 ± 6.56 - - (0.739)
NB$_{cost}$	71.46 ± 5.72 - - (0.287)	58.20 ± 6.78 - - (0.730)
Third-year	4 classes	12 classes
NB	54.26 ± 6.38 - - (0.488)	50.96 ± 6.63 - - (1.685)
NB$_{cost}$	55.02 ± 6.94 - - (0.481)	51.02 ± 6.49 - - (1.645)
Fourth-year	4 classes	15 classes
NB	49.65 ± 7.71 - - (0.540)	50.89 ± 7.38 - - (4.094)
NB$_{cost}$	50.21 ± 7.90 - - (0.539)	49.51 ± 7.35 - - (4.091)

Table V
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

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 year.

IV. Conclusion

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. For this reason, we focus on building junior and senior models to predict the *h-index* of professors of Spanish public universities. These models are learnt from author-based variables (area, position, university, seniority) and 12 bibliometric indices.

The *h-index* prediction is considered as an ordinal classification problem. In this way, we used a cost-sensitive naive Bayes approach which maximizing the classification accuracy, but also minimizing the weighted distances between the actual and the predicted values.

We found 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, our 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, we will use some wrapper feature subset selection and other classification methods, e.g., weighted naive Bayes, tree augmented network, k-nearest neighbour, C4.5, among others. Finally, the *h-index* value could vary depending on the source consulted (Google Scholar, Scopus, ISI WoK, etc.), which is a point to be taken into account.

Acknowledgment

The work has also been partially supported by Spanish Ministry of Science and Innovation, grants TIN2010-20900-C04-04, Cajal Blue Brain and Consolider Ingenio 2010-CSD2007-00018.

References

[1] S. Alonso, F. 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*, vol. 3, no. 4, pp. 273–289, 2009.

[2] L. Egghe, "The hirsch-index are related impact measures," *Annual Review of Information Science and Technology*, vol. 44, pp. 65–114, 2010.

[3] J. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[4] O. Baskurt, "Time series analysis of publication counts of a university: What are the implications?" *Scientometrics*, vol. 86, no. 3, pp. 645–656, 2011.

[5] G. Krampen, A. von Eye, and G. Schui, "Forecasting trends of development of psychology from a bibliometric perspective," *Scientometrics*, vol. 87, no. 2, pp. 687–694, 2011.

[6] A. Ibáñez, P. Larrañaga, and C. Bielza, "Predicting citation count of bioinformatics papers within four years of publication," *Bioinformatics*, vol. 25, no. 24, pp. 3303–3309, 2009.

[7] L. Egghe and R. Rousseau, "An informetric model for the hirsch-index," *Scientometrics*, vol. 69, no. 1, pp. 121–129, 2006.

[8] L. Egghe, "Dynamic h-index: the Hirsch index in function of time," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 3, pp. 452–454, 2006.

[9] F. Ye and R. Rousseau, "The power law model and total career h-index sequences," *Journal of Informetrics*, vol. 2, no. 4, pp. 288–297, 2008.

[10] M. Minsky, "Steps toward artificial intelligence," *IRE*, vol. 49, no. 1, pp. 8–30, 1961.

[11] L. Egghe, "An improvement of the h-index: The g-index," *ISSI Newsletter*, vol. 2, no. 1, pp. 8–9, 2006.

[12] S. Alonso, F. 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*, vol. 82, no. 2, pp. 391–400, 2010.

[13] B. Jin, "h-index: An evaluation indicator proposed by scientist," *Science Focus*, vol. 1, no. 1, pp. 8–9, 2006.

[14] 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*, vol. 59, no. 5, pp. 830–837, 2008.

[15] F. 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*, vol. 4, no. 1, pp. 23–28, 2010.

[16] F. Ruane and R. Tol, "Rational (successive) h-indices: An application to economics in the Republic of Ireland," *Scientometrics*, vol. 75, no. 2, pp. 395–405, 2008.

[17] P. Batista, M. Campiteli, O. Kinouchi, and A. Martinez, "Is it possible to compare researchers with different scientific interests?" *Scientometrics*, vol. 68, no. 1, pp. 179–189, 2006.

[18] J. Soler, "A rational indicator of scientific creativity," *Journal of Informetrics*, vol. 1, no. 2, pp. 123–130, 2007.

[19] A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos, "Generalized hirsch h-index for disclosing latent facts in citation networks," *Scientometrics*, vol. 72, no. 2, pp. 253–280, 2007.

[20] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[21] M. Stone, "Cross-validation choice and assessment of statistical predictions," *Journal of the Royal Statistic Society*, vol. 36, pp. 111–147, 1974.