Genetic Algorithms and Gaussian Bayesian Networks to Uncover the Predictive Core Set of Bibliometric Indices

Alfonso Ibáñez
Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte 28660, Spain. E-mail: fonsoim@gmail.com

Rubén Armañanzas
Krasnow Institute for Advanced Study, George Mason University, 4400 University Drive, Fairfax, VA 22030. E-mail: rarmanan@gmu.edu

Concha Bielza
Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte 28660, Spain. E-mail: mcbielza@fi.upm.es

Pedro Larrañaga
Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte 28660, Spain. E-mail: plarranaga@fi.upm.es

The diversity of bibliometric indices today poses the challenge of exploiting the relationships among them. Our research uncovers the best core set of relevant indices for predicting other bibliometric indices. An added difficulty is to select the role of each variable, that is, which bibliometric indices are predictive variables and which are response variables. This results in a novel multioutput regression problem where the role of each variable (predictor or response) is unknown beforehand. We use Gaussian Bayesian networks to solve the this problem and discover multivariate relationships among bibliometric indices. These networks are learnt by a genetic algorithm that looks for the optimal models that best predict bibliometric data. Results show that the optimal induced Gaussian Bayesian networks corroborate previous relationships between several indices, but also suggest new, previously unreported interactions. An extended analysis of the best model illustrates that a set of 12 bibliometric indices can be accurately predicted using only a smaller predictive core subset composed of citations, g-index, q^2-index, and h_r-index. This research is performed using bibliometric data on Spanish full professors associated with the computer science area.

Received March 19, 2014; revised October 11, 2014; accepted October 24, 2014

 DOI: 10.1002/asi.23467

Introduction

Bibliometric indices are quantitative metrics for assessing the outputs and impacts of individual researchers. They constitute an objective method whose results are reproducible. The main advantage of these indices is that they can summarize the scientific production of an author as a set of figures. This can, at the same time, be a limitation because it removes many details from the citation records. Many funding agencies and promotion committees use these indices regularly as decision-support tools to evaluate research projects and researchers alike. Bibliometric indices are hence an increasingly important topic within the scientific community.

Many studies (Cabezas-Clavijo, Robinson-García, Escabias, & Jiménez-Contreras, 2013; Fu & Aliferis, 2010; Hirsch, 2007; Jensen, Rouquier, & Croissant, 2009; Kissin, 2011; Levitt & Thelwall, 2011; Vieira, Cabral, & Gomez, 2014a, 2014b) have looked at the predictive power of bibliometric indices in many situations: prediction of article impact; scientist promotions; acceptance of grant proposals; and future values of many bibliometric indices, among others. The result is that the scientific community now faces the challenge of selecting which from this pool of bibliometric indices have higher predictive power.

Jensen et al. (2009) found that the h-index was best at predicting promotions within the Centre National de la

Recherche Scientifique (CNRS) researchers. Cabezas-Clavijo et al. (2013) showed that the main bibliometric indicators that explain the funding of Spanish research proposals, in most cases, are the number of papers and the number of papers published in first-quartile journals of the *Journal Citation Reports* (*JCR*). Vieira et al. (2014a, 2014b) 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 article presents a method for identifying a core set of bibliometric indices for prediction purposes. This subset of relevant indices shows a high accuracy when predicting other bibliometric measures. Given a data set of bibliometric indices $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, we tackle the task of selecting which subset best corresponds to predictive variables $\boldsymbol{X}_P$ (variables with a higher predictive power) and which group can be considered as response variables $\boldsymbol{X}_R$, 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 we know the relevant index values (predictive variables), 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, we learn the statistical relationships among the set of bibliometric indices by means of Gaussian Bayesian networks (GBNs) (Geiger & Heckerman, 1994; Shachter & Kenley, 1989). GBNs are specific Bayesian networks whose variables follow a Gaussian distribution. Bayesian networks (BNs) (Heckerman, 1998; Jensen, 2001; Lauritzen, 1996; Pearl, 1988) are graphical models for representing the probabilistic relationships among variables. They consist of two main components: the structure, which is a directed acyclic graph, for representing the dependency and independency among variables (in our case, bibliometric indices), and a set of parameters for representing the quantitative information of the dependency. The learnt GBN is then used to identify how bibliometric indices relate to one another 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 multioutput (Breiman & Friedman, 1997). The goal of multioutput 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 our GBNs, 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 GBN structures based on the distance between real and predicted response variable values. The optimization process searches for the GBN, which minimizes the fitness score. However, the number of possible GBNs is huge, and therefore a genetic algorithm (GA) (Holland, 1975) 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 one another.

The interest and originality of our analysis is a novel multioutput regression problem where the role of each variable (predictor or response) is unknown beforehand. To solve this problem, we introduce a new GBN structure learning algorithm that explores the best GBN structure that minimizes the distance between real and predicted response variable values. The resulting GBN structure reports the most predictive bibliometric indices. From their values, we could calculate accurately the values of bibliometric indices. The scientific community could take advantage of the highly accurate predictions, avoiding the tedious and time-consuming rocess of downloading citation records, organizing the nonstructured data, and computing many bibliometric index values.

The remainder of the article is organized as follows. The section on Related Work presents related work associated with the structural learning of BNs, relationships between bibliometric indices, and the prediction of bibliometric indices. Multioutput Regression and GBNs briefly introduces the definitions of multi-output regression and GBNs. In this section, we also discuss how they relate to each other. Learning GBNs Using GAs describes the different elements of the GA on which the GBN learning process is based. The Results section reports the results of applying our approach to a data set of Spanish full professors of computer science. It covers the data set compilation, the experimental setup, the optimal GBNs and a discussion on the best induced GBN. Finally, the results and conclusions are discussed in the Discussion and Conclusions section.

Related Work

This section reviews the state of the art regarding the three issues covered in this article. First, we list algorithms for BN structure learning. Second, we present some research analyzing the relationships among well-known bibliometric indices. Third, we review different approaches to the prediction of bibliometric indices. Throughout the section, our proposal is compared with the reviewed work.

BN Structure Learning

The most difficult task in BNs is to determine their structure, that is, which node should be connected to which node.

DOI: 10.1002/asi

The task of automatically defining structure from a data set is called BN structure learning. There are two basic approaches to BN structure learning from data: algorithms based on constrained methods and score+search methods.

Constraint-based methods (De Campos, 1998; Margaritis, 2005; Smith & Whittaker, 1998; Spirtes, Glymour, & Scheines, 1993) use conditional independence tests to identify the dependent and independent relationships among variables. A major weakness of these methods is that too many tests may have to be performed, with each test being built upon the results of another. This may lead to compound errors in structure identification. Additionally, increasing cardinality in the conditioning part dramatically reduces test reliability. Thus, most of the 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 function and a search strategy) have to be defined. Once a score metric (Akaike, 1974; Cooper & Herskovits, 1992; Heckerman, Geiger, & Chickering, 1995; Rissanen, 1978; Schwarz, 1978) is specified, a search method is needed to find the structure with the optimal score. The fitness score measures the quality of every candidate structure with respect to a data set. The number of candidate structures that can be built from data grows exponentially as the number of variables increases, so an exhaustive search is not a sensible approach to the problem (Chickering, 1995). Therefore, several search strategies may be used to iterate comparisons on reduced sets of structures. The K2 algorithm (Cooper & Herskovits, 1992) is one of the best known score-based algorithms in BNs. It uses the marginal likelihood of the data set given the structure as the score to greedily learn a BN. This is a very active field of research, and there have been several new proposals (Provan & Singh, 1995; Cheng, Greiner, Kelly, Bell, & Liu, 2002; Blanco, Larrañaga, Inza, & Sierra, 2004; Yehezkel & Lerner, 2009; Vidaurre, Bielza, & Larrañaga, 2010; Bui & Jun, 2012; Huang et al., 2013) in the last years.

Researchers have also tackled the problem of BN structure learning using evolutionary algorithms. Larrañaga, Karshenas, Bielza, and Santana (2012) reviewed how BNs have been used in evolutionary algorithms. Some researchers (Cowie, Oteniya, & Coles, 2007; De Campos, Fernández-Luna, Gámez, & Puerta, 2002; Pinto, Naegele, Dejori, Runkler, & Sousa, 2008; Wu, McCall, & Corne, 2010) analyzed the BN structure learning using evolutionary approaches, such as ant colony optimization and particle swarm optimization. In contrast, most investigators use GAs for the purpose of structure learning. In this case, Larrañaga, Poza, Yurramendi, Murga, and Kuijpers (1996) learnt the BN structure that maximizes the K2 score metric. Etxeberria, Larrañaga, and Pikaza (1997) searched the BN structure that best fits the collected data according to a penalized K2 criterion. Myers, Laskey, and DeJong (1999) extended the use of GAs for BN learning to domains with missing data. The search process was guided by the BDe (Bayesian metric with Dirichlet priors and equivalence) score of each network structure. Van Dijk, Thierens, and van der Gaag (2003) searched the BN structure that best fitted the collected data according to a metrics definition language metric. In contrast, Martínez-Morales, Garza-Domínguez, Cruz-Ramírez, Guerra-Hernández, and Jiménez-Andrade (2004) induced BNs taking advantage of the information provided by a combination of different scores, instead of applying a single one. Other researchers learnt other types of BNs. Tucker, Liu, and Ogden-Swift (2001) used a GA for searching the best dynamic BN structure (Friedman, Murphy, & Russell, 1998). Jia, Liu, and Yu (2005) also applied an immune GA for learning dynamic BN. Mascherini and Stefanini (2005) presented a GA to learn the structure of conditional GBNs (Lauritzen, 1992; Lauritzen & Wermuth, 1989). They used an extension of the BDe metric to measure the fitness of candidate structures.

Unlike most of the research we just described that used discretized values, we tackle the problem of learning GBNs that can deal with continuous values. In contrast to the usual metrics, we minimize the distance between real and predicted response variable values. We make use of GAs for structure learning as in previous works. However, our approach finds the optimal structure by means of a wrapper analysis that outputs information on the core bibliometric indices.

Relationships Between Bibliometric Indices

Many bibliometric indices have been proposed in the literature (see Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009; Egghe, 2010). One of the most successful and best known is the h-index (Hirsch, 2005). This index combines productivity and visibility in a single indicator. Despite its advantages, the h-index has some limitations (Costas & Bordons, 2007): (a) It should not be used to compare researchers from different disciplines; (b) it depends on the duration of each researcher's career; (c) it tends to underestimate the achievement of researchers that have a selective publication strategy; and (d) it cannot distinguish between active and inactive researchers. To overcome the limitations of the h-index, different bibliometric indices have been suggested in the literature (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2010; Batista, Campiteli, Kinouchi, & Martinez, 2006; Bornmann, Mutz, & Daniel, 2008a; Cabrerizo, Alonso, Herrera-Viedma, & Herrera, 2010; Egghe, 2006b; Jin, 2006; Ruane & Tol, 2008; Sidiropoulos, Katsaros, & Manolopoulos, 2007; Soler, 2007).

Some studies (Bollen, Van de Sompel, Hagberg, & Chute, 2009; Hirsch, 2007; Leydesdorff, 2009) have examined correlations between a list of bibliometric indices using the Pearson coefficient. Hirsch (2007) analyzed which bibliometric index (h-index, number of papers, number of citations, and mean citations per paper) was best able to predict the future scientific achievement. He showed correlation coefficients between pairs of bibliometric values in order to test their predictive power. His results indicated that the h-index was the best bibliometric index ($\rho = .89$)

DOI: 10.1002/asi

for this purpose. Bollen et al. (2009) found statistically significant correlations between 39 measures of scholarly impact, although the exact values were not reported. Leydesdorff (2009) also showed high correlations between indices, especially the 5-year impact factor and article influence ($\rho = .956$). Other studies, such as Franceschet (2010), also investigated the degree of correlation between some typical bibliometric indices, tested using the Spearman correlation. This study found strong correlations between the examined indices. The strongest correlation was between the 2-year impact factor and the 5-year impact factor ($\rho_S = .96$). Unlike Hirsch (2007), Bollen et al. (2009), Leydesdorff (2009), and Franceschet (2010), who analyzed simple linear correlations between pairs of indices, Ibáñez, Larrañaga, and Bielza (2011b) learnt a BN model from bibliometric data, which was then used to analyze the relationships among triplets of indices.

Building on this seminal work, we now model relationships among bibliometric indices using GBNs, thanks to which we can work directly with continuous values. The new proposal employs GAs instead of classical network algorithms, such as K2, for structure learning. As an extension, new bibliometric indices are added to these models in order to consider some aspects not previously reported in the literature.

Prediction of Bibliometric Indices

Researchers have addressed the prediction of bibliometric indices, such as the number of documents, the number of citations, or the h-index, among others. For example, Krampen, von Eye, and Schui (2011) forecasted the number of documents that a researcher would produce within 10 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. Ibáñez, Larrañaga, and Bielza (2009) predicted the number of citations of bioinformatics articles within 4 years of publication using tokens found in the abstracts. They used different classification paradigms as predictive models, namely, naïve Bayes, logistic regression, classification trees, and the k-nearest neighbor algorithm. Egghe (2006a) used the power law model to predict the h-index as a function of time, whereas Ye and Rousseau (2008) used nonlinear regression to predict the h-index of authors, journals, and universities. Both studies used former h-index values to extrapolate the h-index value in the near future. Acuna, Allesina, and Kording (2012) predicted h-index values using a linear regression with elastic net regularization. They used former h-index values together with other bibliometric measures as predictive variables to predict future h-index values. Finally, Ibáñez, Larrañaga, and Bielza (2011a, 2014) used cost-sensitive naïve Bayes and cost-sensitive selective naïve Bayes classifiers to predict the h-index of researchers and journals, respectively, from a set of different bibliometric indices.

Unlike the works we just cited that predict only one variable, here we simultaneously predict a set of response variables. Given that all of them are continuous, this is a multioutput regression problem. More interestingly, the role of each variable (predictor or response) is unknown beforehand, so the goal is to discover a core set of bibliometric indices (the predictors) with a higher predictive power in order to forecast the other bibliometric indices (the responses). Once the values of the predictive indices are known, the response index values can be predicted with high accuracy.

MultiOutput Regression and GBNs

This section first introduces the multioutput regression problem, which simultaneously predicts several response variables using the same predictive variables. It also introduces the definition of GBNs. Finally, an approach to learn multioutput regression using GBNs is presented.

MultiOutput Regression

We first formally describe the multioutput regression problem. 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 multioutput 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, that is, as shown by Equation (1):

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{e}, \tag{1}$$

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.

GBNs

Formally, a BN is defined as a pair of elements (G, P). The first, $G = (V(G), A(G))$, is a directed acyclic graph defined by a set of nodes $V(G)$ and a set of arcs among the nodes, $A(G)$. The nodes represent the random variables of the problem, i.e., $V(G) = \{X_1, \ldots, X_n\}$, and the arcs $A(G) \subseteq V(G) \times V(G)$ are the probabilistic conditional dependencies. The second element of every BN, P, represents the joint probability distribution of $(X_1, \ldots, X_n)$ within G, defined as Equation (2):

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \Pi(X_i)), \tag{2}$$

DOI: 10.1002/asi

where $\Pi(X_i)$ represents the set of parents of X_i. A node X_j is a parent of another node X_i if there is an arc from X_j to X_i in G.

A BN is said to be a GBN if, and only if, its associated joint probability distribution is a multivariate normal distribution, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with a joint probability density function (Equation [3]):

$$f(\boldsymbol{x}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad (3)$$

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

The joint probability distribution of the variables in a GBN can be specified as in Equation (2) by the product of a set of conditional probability distributions (Equation [4]):

$$f(x_i \mid \boldsymbol{\pi}(x_i)) \sim \mathcal{N}\left(\mu_i + \sum_{x_j \in \boldsymbol{\Pi}(X_i)} \beta_{ij}(x_j - \mu_j), v_i\right), \quad (4)$$

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 $\boldsymbol{\Pi}(X_i)$, and v_i is the conditional variance of X_i given its parents. It can be calculated as Equation (5):

$$v_i = \Sigma_{X_i} - \boldsymbol{\Sigma}_{X_i\boldsymbol{\Pi}(X_i)} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\Pi}(X_i)} \boldsymbol{\Sigma}^T_{X_i\boldsymbol{\Pi}(X_i)}, \quad (5)$$

where Σ_{X_i} is the unconditional variance of X_i, $\boldsymbol{\Sigma}_{X_i\boldsymbol{\Pi}(X_i)}$ is the row matrix with covariances between X_i and $\boldsymbol{\Pi}(X_i)$, and $\boldsymbol{\Sigma}_{\boldsymbol{\Pi}(X_i)}$ is the covariance matrix of $\boldsymbol{\Pi}(X_i)$. Finally, Figure 1 shows an example of GBN structure and its joint probability distribution.

Learning MultiOutput Regression Using GBNs

The multioutput regression problem can be tackled using a GBN 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 nonevidential variables, we assume a joint multivariate Gaussian distribution with mean vector and covariance matrix given by

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{pmatrix},$$

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 covariance matrix, $\boldsymbol{\Sigma}$, is of great interest in GBNs because its inverse matrix, the precision matrix ($\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$), captures the dependence structure of the variables of the problem. Anderson (2003) demonstrated that a variable X_i is conditionally independent of a variable X_j given the rest of the variables if the value $w_{ij} = 0$. Previously, Shachter and Kenley (1989) found that given the densities of Equation (4), it is possible to determine the precision matrix $\boldsymbol{W}$. They used the following recursive formula (Equation [6]):

$$\boldsymbol{W}(i+1) = \begin{pmatrix} \boldsymbol{W}(i) + \dfrac{\boldsymbol{\beta}_{i+1}\boldsymbol{\beta}^T_{i+1}}{v_{i+1}} & \dfrac{-\boldsymbol{\beta}_{i+1}}{v_{i+1}} \\ \dfrac{-\boldsymbol{\beta}^T_{i+1}}{v_{i+1}} & \dfrac{1}{v_{i+1}} \end{pmatrix}, \quad (6)$$

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 the variables given some observations. Here, we follow a method presented by Castillo, Gutiérrez, and Hadi (1997) to perform evidence propagation in a GBN. Given this joint distribution, the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is multivariate Gaussian with mean vector $\boldsymbol{\mu}^{Y|X=x}$ and covariance matrix $\boldsymbol{\Sigma}^{Y|X=x}$ given by Equations (7) and (8):

$$\boldsymbol{\mu}^{Y|X=x} = \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}^{-1}_{XX}(\boldsymbol{x} - \boldsymbol{\mu}_X), \quad (7)$$

$$\boldsymbol{\Sigma}^{Y|X=x} = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}^{-1}_{XX}\boldsymbol{\Sigma}_{XY}. \quad (8)$$

Finally, we note that Equations (1) and (7), and (8) are different parameterizations of the same regression model, given that $\boldsymbol{B} = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}^{-1}_{XX}$.

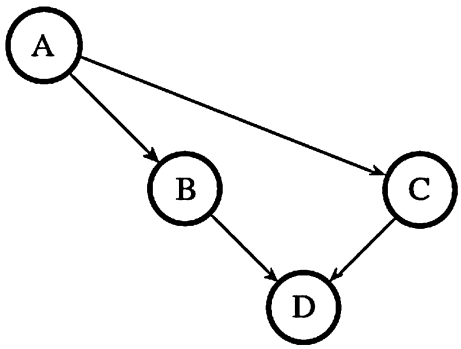

FIG. 1. GBN structure and its joint probability distribution.

Learning GBNs Using GAs

GAs 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. GAs try to find better individuals (solutions for the given problem) by producing fitter descendants in a set of populations.

GAs and GBNs are used in this article to uncover the subset of bibliometric indices with the highest predictive power of all. A wrapper analysis to evaluate all different

DOI: 10.1002/asi

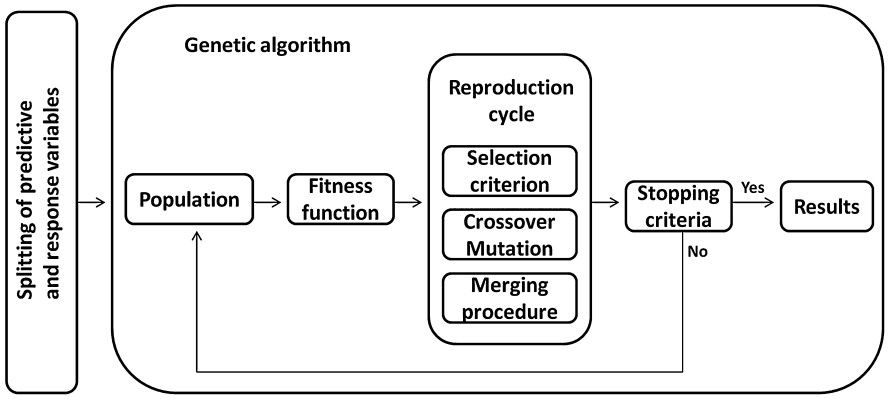

FIG. 2. Steps of our genetic algorithm method.

structures is used to accomplish this goal. Therefore, after setting up a specific splitting of predictive and response variables, we use a GA to search the optimal GBN 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 2 shows our GA methodology.

Details of our implementation, such as individual codification, fitness function, selection, crossover, mutation, and termination criterion, follow.

Initial Population

The search space of candidate solutions is represented as a collection of N individuals, called population. In our problem, individuals represent GBN 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 string $(a_{11}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, \ldots, a_{n1}, \ldots, a_{nn})$, 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 $p = .5$ (probability of success). If need be, the network structure is amended to avoid the presence of cycles.

Fitness Function

The calculation of the fitness function accounts for the main computational burden in a GA. An ideal fitness function should correlate closely with the goal and should be computed quickly. This running time is crucial given that the GA has to be iterated several times to produce reliable results in nontrivial problems.

In our study, given an individual with p predictive variables and r response variables, we calculate the Mahalanobis distances between real and predicted values for the r response variables (see Equation [9]). 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:

$$MD(\boldsymbol{y}, \boldsymbol{y}') = \sqrt{(\boldsymbol{y} - \boldsymbol{y}')^T \Sigma_{YY}^{-1} (\boldsymbol{y} - \boldsymbol{y}')}, \tag{9}$$

where $\boldsymbol{y}$ and $\boldsymbol{y}'$ are vectors representing the real and predicted values of the response variables and Σ_{YY} is the covariance matrix of the response variables.

We use the Mahalanobis distance 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), we select a distance-based score because our objective is to minimize the distance between real and predicted response variable values. We decided to use the Mahalanobis distance, instead of the Euclidean distance, because it has some advantages, that is, 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.

Reproduction Cycle

Parent selection criterion, crossover, and mutation operators and merging procedure are the constituents of the reproduction cycle in a GA. We explore the details of each part here.

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, such as proportional selection methods, ranking selection, tournament selection, and so on, are available in the literature (Sivaraj & Ravichandran, 2011). We use an elitist strategy,

DOI: 10.1002/asi

which chooses the best k individuals from the population as parents for reproduction. This strategy guarantees the improvement of the average and minimum value in each GA 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 (Spears & Anand, 1991). We choose the single-point crossover operator, which is the most common operator and provides good results (Kellegoza, Toklub, & Wilsonc, 2008). Given the binary codification of the individuals, we randomly choose with a fixed probability P_c a crossover point 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. In our research, we use single-point mutation: 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, we choose an elitist strategy 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.

Stopping Criteria

The search is halted when a set of conditions, the stopping criteria, are satisfied. Different criteria for stopping a GA 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 our stopping criteria. The individual with the highest score in the final population is considered to be the solution to the optimization problem.

Results

Data Set Compilation

The first stage of the data collection process was to contact the Spanish Ministry of Education to request the list of full professors of computer science who were active as of January 1, 2010. This list includes 280 faculty members with their full names and affiliations. For each faculty member, we compiled a list of publications and citation data from 1973 (the first year with records) to 2010. The data were retrieved from the Thomson Reuters Web of Knowledge (WoK). WoK contains databases specialized in journals, such as *Science Citation Index* and *JCR*, and in conferences such as *Conference Proceedings Citation Index*. In total, WoK indexes more than 470 computer science journals and more than 15,000 of the major computer science conferences.

Although the WoK does not store all the scientific literature, it does record a very large part (Garfield, 1996). A careful curation of the data is needed owing to problems related to Spanish personal name variations in international databases (Ruiz-Pérez, Delgado-López-Cózar, & Jiménez-Contreras, 2002). The records were also filtered by the publication subject. Only documents published in journals and conferences belonging to the seven major fields of computer science were finally included. According to the *JCR* rankings, these major fields are: artificial intelligence; cybernetics; hardware and architecture; information systems; interdisciplinary applications; software engineering; and theory and methods. To ensure data reliability, we checked our final list of publications against other databases such as the DBLP Computer Science Bibliography, personal web pages, and institutional websites, among others. Finally, a list of bibliometric indices (documents, citations, h-index [Hirsch, 2005], g-index [Egghe, 2006b], hg-index [Alonso et al., 2010], a-index [Jin, 2006], m-index [Bornmann et al., 2008a], q^2-index [Cabrerizo et al., 2010], h_r-index [Ruane & Tol, 2008], h_i-index [Batista et al., 2006], h_c-index [Sidiropoulos et al., 2007], and c-index [Soler, 2007]) were computed for each academic within the database. A short description of each index follows.

X_1 *(documents).* Associated with the number of published articles, it represents the output of each professor.

X_2 *(citations).* Total number of citations received by the publication portfolio of a researcher, it represents a measure of research visibility.

X_3 *(h-index).* The h-index quantifies the scientific output of a single researcher as a single-number criterion. It is based on a list of publications ranked in descending order of the number of citations. The value of the h-index is equal to the number of articles (h) in the list that have h or more citations. The h-index incorporates both the quantity and the visibility of a publication portfolio, although not always in a consistent manner.

DOI: 10.1002/asi

X_4 *(g-index).* The h-index tends to underestimate the achievement of researchers who have a selective publication strategy. This strategy is followed by researchers who publish fewer documents than average, but, by contrast, receive many citations. To avoid such bias, the g-index 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 articles, favoring researchers with a selective publication strategy. Although the g-index is better than the h-index in this sense, is not a fully satisfactory solution.

X_5 *(hg-index).* The hg-index is a combination of both the h-index and g-index. It aims to provide a more balanced view of scientific production. The hg-index of a researcher is defined 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, respectively.

X_6 *(a-index).* The a-index is defined as the average number of citations received by the articles included in the h-core, that is, the first h articles. This index measures the citation intensity of the h-core articles; however, it can be very sensitive to just a few articles receiving high citation counts.

X_7 *(m-index).* The distribution of citation counts is usually skewed; hence, the median is a better measure of central tendency. The m-index computes the median number of citations received by articles in the h-core.

X_8 *(q^2-index).* This index provides a more global view of scientific production. It is based on the geometric mean of the h-index, describing the number of the articles (quantitative dimension), and the m-index, depicting the impact of the articles (qualitative dimension)

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

where m corresponds to the value of the m-index.

X_9 *(h_r-index).* The rational h-index is an extension of the original h-index. It reflects the number of citations needed to increase the h-index by one unit. Mathematically,

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

where $Cit(h+1)$ is the number of citations received by the $(h+1)$-th article.

X_{10} *(h_i-index).* The individual h-index is complementary to the h-index and estimates the number of articles that a researcher would have written throughout his career with at least h_i citations if he had worked alone. The rationale behind this is to measure the effective individual average productivity:

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

where N_a is the mean number of authors in the h-core articles.

X_{11} *(h_c-index).* The original h-index cannot distinguish between inactive scientists, junior scientists, and senior scientists. To account for this temporal component, a score $Sc(i)$ was defined for an article 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 article i; $Cit(i)$ is the total number of citations received by article i; γ and δ are arbitrary parameters. Using this score, the value of old articles gradually declines, even if they still receive citations. Therefore, the definition of the contemporary h-index states that "a researcher has index h_c, if h_c of his published papers get a score of $Sc(i) \geq h_c$ each, and the other papers get a score of $Sc(i) < h_c$".

X_{12} *(c-index).* This index measures creativity, defined as the generation of new scientific knowledge. Its purpose is to highlight articles that receive many citations and have few bibliographic references. This index is calculated from the list of citations and references of the author's articles:

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

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

We have selected this small subset of well-known indices to provide a practical example using our method. 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 purpose given that most of them are size-dependent indicators, which sometimes behave in a counterintuitive manner (Marchant, 2009; Waltman & van Eck, 2012). In this way, there are better bibliometric indicators, such as 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. It is not our aim to argue in favor of the selected indicators as good ones to assess scientists; we select

DOI: 10.1002/asi

TABLE 1. Statistical figures of all bibliometric indices computed from the publications data set of 280 active Spanish full professors of computer science (years 1973–2010).

Variables	Min	First quartile	Mean	Median	Third 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 2. Correlation coefficients among bibliometric indices computed from the publications data set of 280 active Spanish full professors.

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

Note. 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), and X_{12} (c-index).

them as variables for our GBN models. Given these variables, our 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.

In order to give an overview of indices values, Table 1 shows a statistical summary for each index. Note that the average academic publishes 34.8 documents and receives 143.0 citation. Also noticeable is that the average h-index value is 6. Citation values range from 1 to 4,570 citations during the period, that is, 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.

Finally, Table 2 shows the correlation coefficients among bibliometric indices. Most of the selected bibliometric indices are variants, extensions, or generalizations of the h-index. This implies that these indices are usually correlated among themselves, which is corroborated by the mean correlation coefficient ($\rho = .82$) among selected indices. We note that documents is the index that has the lowest correlations, that is, there are weak correlations between documents and m-index ($\rho = .45$), documents and a-index ($\rho = .52$), and documents and c-index ($\rho = .61$), among others. In contrast, the hg-index has the highest correlations, that is, there are strong correlations between hg-index and h-index ($\rho = .99$), hg-index and g-index ($\rho = .99$), and hg-index and h_r-index ($\rho = .99$), among others.

Experimental Setup

The application of a GA means setting several parameters such as 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 that work

DOI: 10.1002/asi

consistently well for function optimization (De Jong & Spears, 1990; Grefenstette, 1986). In this article, we follow Grefenstette's recommendations with minor changes.

According to Grefenstette (1986), the population size should be 30 individuals. We reduce the number of individuals to 20 because our fitness function is time-consuming to compute. Even so, our population is sufficient to uncover the subset of bibliometric indices with the highest predictive power. Crossover probability is .9, whereas mutation probability is set at .01. We use a single-point coupled crossover operator and a single-point mutation operator. The algorithm halts after reaching 40 generations or when there is no improvement after five consecutive generations.

In our study, all different structures with p predictive variables and r response variables are explored. Because the data set includes 12 bibliometric indices, there is a total of 4,096 ($= 2^{12}$) different splittings. Once the role of predictive and response nodes has been fixed, the GA 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.

In order to have a fair performance estimation, we choose k-fold cross-validation as the procedure for estimating the predictive accuracy. This method divides all cases from the data set 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 fitness scores are then averaged to output the actual estimation (Stone, 1974). In our experiments, we use a value of 5 for k in the cross-validation procedure.

Optimal GBNs

The result of our GA is a set of 11 optimal GBNs. Each model is associated with a different cardinality of predictive and response variables, that is, one predictive variable and 11 response variables, two predictive variables and 10 response variables, and so on. To assess the improvement produced by each optimal network, we first define two general and specific baseline values for comparison. Both of these baselines correspond to the Mahalanobis distances between predicted and real values of the response variables when naïve network structures are considered, that is, networks without arcs between nodes.

The general baseline corresponds to the average Mahalanobis distance of all naïve structures with the same number of response variables, regardless of which variables they are. Conversely, the specific baseline accounts for the Mahalanobis distance of a naïve structure using the same response variables as the network used for comparison. The rationale behind this is to confirm that our optimal GBNs are better than both general and specific baselines.

Table 3 shows the list of predictive variables of our 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, we observe 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 naïve model with the same number of response variables (3.013) and with the same splitting of variables (2.931).

TABLE 3. Predictive variables for the identified optimal GBNs: 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

Note. 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), and X_{12} (c-index).

It is also of interest to compare the performance of the best GBNs with one another. To do so, we compute the fitness of the models given the data. Classical *goodness-of-fit* criteria rank complex models, that is, models with more parameters, 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 trade-off between model complexity and goodness-of-fit is the Bayesian information criterion or BIC (Schwarz, 1978). BIC penalizes the complexity of a model by an additional term, depending on the number of parameters of the model and the sample size. This way BIC provides a quantitative measure for model selection. We select the model with the highest BIC value.

Table 4 collects the BIC score of each optimal GBN. The highest BIC value of all (−6,574.755) 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]). The next section details its full structure, conditional dependencies, and predictive performance.

Discussion of the Best Induced GBN

The network that 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.

DOI: 10.1002/asi

TABLE 4. Predictive variables for the identified optimal GBNs: BIC values for the reported model.

No. of predictors	Predictive variables	BIC values
1	X_9	–7,783.949
2	X_3, X_4	–7,028.680
3	X_3, X_4, X_{11}	–7,020.569
4	X_2, X_4, X_8, X_9	–6,574.755
5	X_2, X_5, X_6, X_9, X_{12}	–7,106.992
6	X_1, X_2, X_5, X_8, X_9, X_{12}	–6,816.705
7	X_2, X_4, X_6, X_8, X_9, X_{10}, X_{12}	–6,871.926
8	X_1, X_2, X_3, X_5, X_6, X_7, X_{10}, X_{12}	–6,933.521
9	X_1, X_2, X_3, X_5, X_6, X_7, X_{10}, X_{11}, X_{12}	–6,655.728
10	X_1, X_2, X_3, X_5, X_6, X_7, X_8, X_{10}, X_{11}, X_{12}	–7,232.618
11	X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_{10}, X_{11}, X_{12}	–8,037.581

Note. 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), and X_{12} (c-index).

Network structure. Figure 3 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.

We examine a set of centrality measures in order to analyze the graphical characteristics of the network in Figure 3. 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. We define two separate measures of centrality degree: indegree and outdegree. A node's indegree is the number of arcs directed to the node, and 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: $CD(X_1) = 5$; $CD(X_2) = 5$; $CD(X_3) = 7$; $CD(X_4) = 10$; $CD(X_5) = 8$; $CD(X_6) = 6$; $CD(X_7) = 7$; $CD(X_8) = 7$; $CD(X_9) = 8$; $CD(X_{10}) = 6$; $CD(X_{11}) = 7$; 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 the potential relationship between strong correlation coefficients and network structure, we observe that strong correlations are not a problem for the method presented. A potential high correlation coefficient does not imply an arc among correlated variables in our GBN. In this way, Table 2 shows strong correlations between the hg-index and h_r-index ($\rho = .99$) and between the h-index and h_i-index ($\rho = .93$), which are not presented as arcs in Figure 3. In contrast, the weak correlation between documents and the m-index ($\rho = .45$) is presented as an arc in Figure 3. The presence of arcs does not depend on potential strong correlations; it depends on our GA, which looks for the optimal structure that minimizes the distance between real and predicted response variable values.

Dependencies among indices. Based on the definitions of the indices (see earlier section on Data Set Compilation), it is clear that some of them can be expressed according to the values of other indices. For example, the hg-index could be expressed in terms of h- and g-index values. Also, the 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 the hg-index (X_5) in the network structure of Figure 3, and the h-index (X_3) and the m-index (X_7) are children of the q^2-index (X_8).

Besides revealing dependencies already present in the index definitions, the GBN discovers dependencies that 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, we note that the information about h_r-index influences the density function of the h-index, as expected in the h_r-index, an extension of h-index.

The arc between the a-index and m-index, ($X_6 \rightarrow X_7$) in Figure 3 is an example of a dependency that is initially expected. Remember that the 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 the h-index ($X_1 \rightarrow X_3$), citations and the 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 level of correlation (Bornmann, Wallon, & Ledin, 2008b; Costas & Bordons, 2008; Schreiber, 2008). There are other network dependencies, for example, 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 pointed out similar correlations (Franceschet, 2009).

Conversely, the network included some unexpected arcs. In this way, the GBN reported probabilistic dependencies between the a-index and the h_i-index ($X_6 \rightarrow X_{10}$), the m-index and the h_c-index ($X_7 \rightarrow X_{11}$), and the q^2-index and the c-index ($X_8 \rightarrow X_{12}$).

Conditional independencies among indices. GBNs are a powerful tool not only for capturing dependencies, but also for identifying conditional independencies among variables. Here, we address Markov network properties with the aim of discovering such independencies among the nodes of the best induced network. The local Markov property states that any node X_i in a BN is conditionally independent of its nondescendants given the values of its parents. It can be

DOI: 10.1002/asi

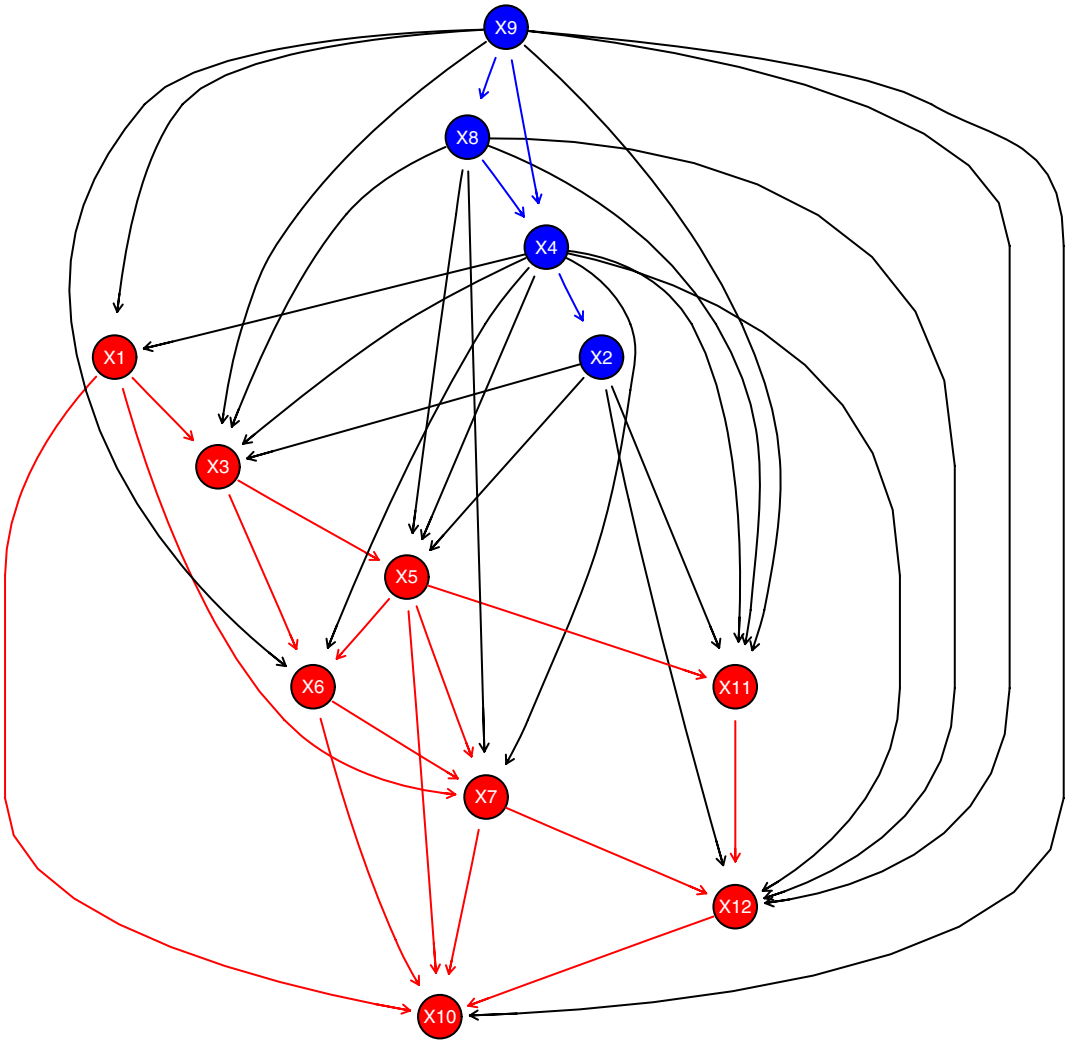

FIG. 3. Best GBN structure. Each node represents: X_1 (documents); X_2 (citations); X_3 (h-index); X_4 (g-index); X_5 (hg-index); X_6 (a-index); X_7 (m-index); X_8 (q^2-index); X_9 (h_r-index); X_{10} (h_i-index); X_{11} (h_c-index); and X_{12} (c-index). Blue nodes correspond to predictive variables ($\boldsymbol{X_P}$), whereas red nodes correspond to response variables ($\boldsymbol{X_R}$). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

expressed as $I(X_i, nondescendants(X_i) \mid \Pi(X_i))$. 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 MB of a node includes its parents, its children, and its children's parents. Thus, $I(X_i, non\text{-}MB(X_i) \mid MB(X_i))$.

Table 5 lists conditional independencies between the bibliometric indices of the network in Figure 3. The list is derived from the local and global Markov properties. This network identifies conditional independencies in accord with index definitions, as well as other conditional independencies that are hidden in such definitions. Taking the hg-index as an example, we find that, given citations, h-index, g-index, and q^2-index, the hg-index is independent of documents and the h_r-index. This suggests that when we know the values of citations, the h-index, g-index, and q^2-index, the value of documents provides no information on the value of the hg-index. Focusing on the q^2-index, we note that it is conditionally independent of the 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 5.

However, there are other conditional independencies that 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, our model shows that the a-index is conditionally independent of documents and citations, given the 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 our GBN indicate a probabilistic, not a causal, relationship.

Predicting bibliometric indices. Now, we inspect the probabilistic component of the network in Figure 3 and what the effect of knowing the values of some variables has on the

DOI: 10.1002/asi

TABLE 5. Conditional independencies among bibliometric indices derived using local and global Markov properties in the GBN of Figure 3.

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

TABLE 6. Evidence propagation results using the GBN of Figure 3 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

others. In doing so, we use evidence propagation 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 GBN 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 6 presents three inference examples. It shows the evidence values for the predictive variables and the predictions made by the network in Figure 3 for the response variables. These predictions are the mean vector ($\boldsymbol{\mu}^{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$) of the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$, which is computed with Equation (7). 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, we fix citations = 853, g-index = 28, q^2-index = 19.4, and h_r-index = 15.9. After setting the evidence, we compute the predicted values of the response indices. Results in Table 6 show that predicted values are very close to real values. Regarding documents, h-index, and hg-index, we observe 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. Last, 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.

Discussion and Conclusions

Bibliometric indices are presently an increasingly important topic for the scientific community. 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 from this pool of bibliometric indices those that have higher predictive power.

A review of the literature presents some recent works (Vieira et al., 2014a, 2014b) that analyzed the success of models based on bibliometric indices in predicting the rankings of applicants to academic positions at the university. As with our 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 our multioutput regression

DOI: 10.1002/asi

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 these studies, we present a novel method 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 behave in a counterintuitive way because of the inconsistencies associated with the mechanism used to aggregate publication and citation statistics into a single number. This article does not argue in favor of the selected indicators as the best bibliometric indices to evaluate research performance; we select them as an example of GBN variables in order to give a practical example using the proposed method.

Given a data set of bibliometric indices, we tackle the task of selecting which subset best corresponds 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 multioutput regression. This results in a novel multioutput regression problem where the role of each variable (predictor or response) is unknown beforehand.

Having split predictive and response variables, we learn a GBN structure to identify relationships among bibliometric indices and for prediction purposes. GBN structure learning is based on a GA, 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 BIC value among the 11 optimal models with different cardinality. Although we conducted an exhaustive analysis to evaluate all possible configurations of predictive and response variables in order to identify the relevant predictive core set of bibliometric indices, a GA could be used for exploring the search domain of different configurations of predictive and response variables.

In our specific problem with full professors, our findings provide information on which subset of bibliometric indices has the highest predictive power. We observe that the bibliometric core is composed of citations, the g-index, the q^2-index, and the h_r-index. This means that when we know the values of these bibliometric indices, the values of the other eight indices can be predicted with a high degree of accuracy. Analyzing its structure, we notice 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 method 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 data set. In this way, the method can be applied to any data set. Obviously, the results usually depend on the selected data set. Despite this, we believe that similar bibliometric indices relationships could be also learnt using different data sets.

In the future, we intend to build alternative models using different BN induction algorithms. It would also be worthwhile to extend the domain of our data collection to overseas researchers and professors alike. These new models could also incorporate other bibliometric indices in order to cover a larger part of the bibliometric domain.

Acknowledgments

Research partially supported by the Spanish Ministry of Economy and Competitiveness (grant no. TIN2013-41592-P) and the Cajal Blue Brain Project (Spanish partner of the Blue Brain Project initiative from EPFL). R.A. is currently supported by grant R01 NS39600 from the National Institutes of Health (NINDS).

References

Acuna, D.E., Allesina, S., & Kording, K.P. (2012). Future impact: Predicting scientific success. Nature, 489(7415), 201–202.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., & Herrera, F. (2009). *h*-index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics, 3(4), 273–289.

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., & Herrera, F. (2010). *hg*-index: A new index to characterize the scientific output of researchers based on the *h*- and *g*-indices. Scientometrics, 82(2), 391–400.

Anderson, T.W. (2003). An introduction to multivariate statistical analysis. New York, USA: Wiley-Interscience.

Batista, P.D., Campiteli, M.G., Kinouchi, O., & Martinez, A.S. (2006). Is it possible to compare researchers with different scientific interests? Scientometrics, 68(1), 179–189.

Blanco, R., Larrañaga, P., Inza, I., & Sierra, B. (2004). Gene selection for cancer classification using wrapper approaches. International Journal of Pattern Recognition and Artificial Intelligence, 18(8), 1373–1380.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. PLoS ONE, 4(6), e6022.

Bornmann, L., Mutz, R., & Daniel, H. (2008a). 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.

Bornmann, L., Wallon, G., & Ledin, A. (2008b). 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.

Breiman, L., & Friedman, J.H. (1997). Predicting multivariate responses in multiple linear regression. Journal of the Royal Statistical Society: Series B (Methodological), 59(1), 3–54.

Bui, A.T., & Jun, C.H. (2012). Learning Bayesian network structure using Markov blanket decomposition. Pattern Recognition Letters, 33(16), 2134–2140.

Cabezas-Clavijo, A., Robinson-García, N., Escabias, M., & Jiménez-Contreras, E. (2013). Reviewers' ratings and bibliometric indicators: Hand in hand when assessing over research proposals? PLoS ONE, 8(6), e68258.

Cabrerizo, F.J., Alonso, S., Herrera-Viedma, E., & Herrera, F. (2010). 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.

Castillo, E., Gutiérrez, J.M., & Hadi, A.S. (1997). Expert systems and probabilistic network models. New York, USA: Springer-Verlag.

Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W.R. (2002). Learning Bayesian networks from data: An information-theory based approach. Artificial Intelligence, 137(1–2), 43–90.

Chickering, D.M. (1995). Learning Bayesian networks is NP-complete. In Proceedings on Artificial Intelligence and statIstics (pp. 121–130). New York, USA: Springer.

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9(4), 309–347.

Costas, R., & Bordons, M. (2007). The *h*-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of Informetrics, 1(3), 193–203.

Costas, R., & Bordons, M. (2008). Is *g*-index better than *h*-index? An exploratory study at the individual level. Scientometrics, 77(2), 267–288.

Cowie, J., Oteniya, L., & Coles, R. (2007). Particle swarm optimization for learning Bayesian networks. In Proceedings of the World Congress on Engineering (pp. 2–4). Hong Kong: Newswood Limited.

De Campos, L. (1998). Independency relationships and learning algorithms for singly connected networks. Journal of Experimental and Theoretical Artificial Intelligence, 10(4), 511–549.

De Campos, L., Fernández-Luna, J.M., Gámez, J.A., & Puerta, J.M. (2002). Ant colony optimization for learning Bayesian networks. International Journal of Approximate Reasoning, 31(3), 291–311.

De Jong, K.A., & Spears, W.M. (1990). An analysis of the interacting roles of population size and crossover in genetic algorithms. In Proceedings of the First Workshop on Parallel Problem Solving from Nature (pp. 38–47). London, UK: Springer.

Egghe, L. (2006a). Dynamic *h*-index: The Hirsch index in function of time. Journal of the American Society for Information Science and Technology, 58(3), 452–454.

Egghe, L. (2006b). An improvement of the *h*-index: The *g*-index. ISSI Newsletter, 2(1), 8–9.

Egghe, L. (2010). The Hirsch-index and related impact measures. Annual Review of Information Science and Technology, 44(1), 65–114.

Etxeberria, R., Larrañaga, P., & Pikaza, J.M. (1997). Analysis of the behaviour of genetic algorithms when learning Bayesian network structure from data. Pattern Recognition Letters, 18(11–13), 1269–1273.

Franceschet, M. (2009). A cluster analysis of scholar and journal bibliometric indicators. Journal of the American Society for Information Science and Technology, 60(10), 1950–1964.

Franceschet, M. (2010). Journal influence factors. Journal of Informetrics, 4(3), 239–248.

Friedman, N., Murphy, K., & Russell, S. (1998). Learning the structure of dynamic probabilistic networks. In Proceedings of the 14th Conference Conference on Uncertainty in Artificial Intelligence (pp. 139–147). San Francisco, USA: Morgan Kaufmann.

Fu, L.D., & Aliferis, C.F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. Scientometrics, 85(1), 257–270.

Garfield, E. (1996). The significant scientific literature appears in a small core of journals. Scientist, 10(17), 13–16.

Geiger, D., & Heckerman, D. (1994). Learning Gaussian networks. In Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence (pp. 235–243). San Francisco, USA: Morgan Kaufmann.

Grefenstette, J.J. (1986). Optimization of control parameters for genetic algorithms. IEEE Transactions on Systems, Man, and Cybernetics, 16(1), 122–128.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. Cambridge, MA: MIT Press.

Heckerman, D., Geiger, D., & Chickering, D.M. (1995). Learning Bayesian networks. The combination of knowledge and statistical data. Machine Learning, 20(3), 197–243.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, 102(46), 16569–16572.

Hirsch, J.E. (2007). Does the h index have predictive power? Proceedings of the National Academy of Sciences, 104(49), 19193–19198.

Holland, J.H. (1975). Adaptation in natural and artificial systems. Ann Arbor, USA: University of Michigan Press.

Huang, S., Li, J., Ye, J., Fleisher, A., Chen, K., Wu, T., . . . Alzheimer's Disease Neuroimaging Initiative. (2013). 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.

Ibáñez, A., Larrañaga, P., & Bielza, C. (2009). Predicting citation count of Bioinformatics papers within four years of publication. Bioinformatics (Oxford, England), 25(24), 3303–3309.

Ibáñez, A., Larrañaga, P., & Bielza, C. (2011a). Predicting the *h*-index with cost-sensitive naive Bayes. In Proceedings of the 11th International Conference on Intelligent Systems Design and Applications (pp. 599–604). Piscataway, USA: IEEE Press.

Ibáñez, A., Larrañaga, P., & Bielza, C. (2011b). Using Bayesian networks to discover relationships between bibliometric indices. A case study of computer science and artificial intelligence journals. Scientometrics, 89(2), 523–551.

Ibáñez, A., Bielza, C., & Larrañaga, P. (2014). Cost-sensitive selective naive Bayes classifiers for predicting the increase of the *h*-index for scientific journals. Neurocomputing, 135, 42–52.

Jensen, F.V. (2001). Bayesian networks and decision graphs. New York, USA: Springer.

Jensen, P., Rouquier, J.B., & Croissant, Y. (2009). Testing bibliometric indicators by their prediction of scientists promotions. Scientometrics, 78(3), 467–479.

Jia, H.Y., Liu, D.Y., & Yu, P. (2005). Learning dynamic Bayesian network with immune evolutionary algorithm. In Proceedings of 2005 International Conference on Machine Learning and Cybernetics (pp. 2934–2938). Piscataway, USA: IEEE Press.

Jin, B. (2006). *h*-index: An evaluation indicator proposed by scientist. Science Focus, 1(1), 8–9.

Kellegoza, T., Toklub, B., & Wilsonc, J. (2008). Comparing efficiencies of genetic crossover operators for one machine total weighted tardiness problem. Applied Mathematics and Computation, 199(2), 590–598.

Kissin, I. (2011). Can a bibliometric indicator predict the success of an analgesic? Scientometrics, 86(3), 785–795.

Krampen, G., von Eye, A., & Schui, G. (2011). Forecasting trends of development of psychology from a bibliometric perspective. Scientometrics, 87(2), 687–694.

Larrañaga, P., Poza, M., Yurramendi, Y., Murga, R.H., & Kuijpers, C.M.H. (1996). 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.

Larrañaga, P., Karshenas, H., Bielza, C., & Santana, R. (2012). A review on probabilistic graphical models in evolutionary computation. Journal of Heuristics, 18(5), 795–819.

Lauritzen, S.L. (1992). Propagation of probabilities, means, and variances in mixed graphical association models. Journal of the American Statistical Association, 87(420), 1098–1108.

Lauritzen, S.L. (1996). Graphical models. Oxford, UK: Clarendon Press.

Lauritzen, S.L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. Annals of Statistics, 17(1), 31–57.

Levitt, J.M., & Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. Information Processing & Management, 47(2), 300–308.

Leydesdorff, L. (2009). 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.

Marchant, T. (2009). Score-based bibliometric rankings of authors. Journal of the American Society for Information Science and Technology, 60(6), 1132–1137.

Margaritis, D. (2005). Distribution-free learning of Bayesian network structure in continuous domains. In Proceedings of the 20th National Conference on Artificial Intelligence (pp. 825–830). Palo Alto, USA: AAAI Press.

Martínez-Morales, M., Garza-Domínguez, R., Cruz-Ramírez, N., Guerra-Hernández, A., & Jiménez-Andrade, J.L. (2004). A method based on genetic algorithms and fuzzy logic to induce Bayesian networks. In Proceedings of the Fifth Mexican International Conference in Computer Science (pp. 176–180). Piscataway, USA: IEEE Press.

Mascherini, M., & Stefanini, F. (2005). M-GA: A genetic algorithm to search for the best conditional Gaussian Bayesian network. In Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (pp. 61–67). Piscataway, USA: IEEE Press.

Myers, J.W., Laskey, K.B., & DeJong, K.A. (1999). Learning Bayesian networks from incomplete data using evolutionary algorithms. In 15th Conference on Uncertainty in Artificial Intelligence (pp. 476–485). San Francisco, USA: Morgan Kaufmann.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Francisco, USA: Morgan Kaufmann.

Pinto, P.C., Naegele, A., Dejori, M., Runkler, T.A., & Sousa, J.M.C. (2008). Learning of Bayesian networks by a local discovery ant colony algorithm. In IEEE Congress on Evolutionary Computation (pp. 2741–2748). Piscataway, USA: IEEE Press.

Provan, G.M., & Singh, M. (1995). Learning Bayesian networks using feature selection. In Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics (pp. 450–456). New York, USA: Springer.

Rissanen, J. (1978). Modeling by shortest data description. Automatica, 14(5), 465–471.

Ruane, F., & Tol, R. (2008). Rational (successive) *h*-indices: An application to economics in the Republic of Ireland. Scientometrics, 75(2), 395–405.

Ruiz-Pérez, R., Delgado-López-Cózar, E., & Jiménez-Contreras, E. (2002). Spanish personal name variations in national and international biomedical databases: Implications for information retrieval and bibliometric studies. Journal of the Medical Library Association, 90(4), 411–430.

Schreiber, M. (2008). 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.

Schwarz, G. (1978). Estimating the dimensions of a model. Annals of Statistics, 6(2), 461–464.

Shachter, R.D., & Kenley, C.R. (1989). Gaussian influence diagrams. Management Science, 35(5), 527–550.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized Hirsch *h*-index for disclosing latent facts in citation networks. Scientometrics, 72(2), 253–280.

Sivaraj, R., & Ravichandran, T. (2011). A review of selection methods in genetic algorithms. International Journal of Engineering Science and Technology, 3(5), 3792–3797.

Smith, P.W.F., & Whittaker, J. (1998). Edge exclusion tests for graphical Gaussian models. In Proceedings of the NATO Advanced Study Institute on Learning in Graphical Models (pp. 555–574). MA, USA: Kluwer Academic Publishers Norwell.

Soler, J.M. (2007). A rational indicator of scientific creativity. Journal of Informetrics, 1(2), 123–130.

Spears, W.M., & Anand, V. (1991). A study of crossover operators in genetic programming. In Proceedings of the Sixth International Symposium on Methodologies for Intelligent Systems (pp. 409–418). London, UK: Springer.

Spirtes, P., Glymour, C., & Scheines, R. (1993). Causation, prediction and search. New York, USA: Springer.

Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. Journal of the Royal Statistic Society, 36(2), 111–147.

Tucker, A., Liu, X., & Ogden-Swift, A. (2001). Evolutionary learning of dynamic probabilistic models with large time lags. International Journal of Intelligent Systems, 16(5), 621–645.

Van Dijk, S., Thierens, D., & van der Gaag, L. (2003). Building a GA from design principles for learning Bayesian networks. In Fifth Annual Conference on Genetic and Evolutionary Computation (pp. 886–897). Heidelberg, Germany: Springer.

Vidaurre, D., Bielza, C., & Larrañaga, P. (2010). 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.

Vieira, E.S., Cabral, J.A.S., & Gomez, J.A.N.F. (2014a). 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.

Vieira, E.S., Cabral, J.A.S., & Gomez, J.A.N.F. (2014b). How good is a model based on bibliometric indicators in predicting the final decisions made by peers? Journal of Informetrics, 8(2), 390–405.

Waltman, L., & van Eck, N.J. (2012). The inconsistency of the *h*-index. Journal of the American Society for Information Science and Technology, 63(2), 406–415.

Wu, Y., McCall, J., & Corne, D. (2010). Two novel ant colony optimization approaches for Bayesian network structure learning. In Proceedings of the 2010 World Congress on Computational Intelligence (pp. 4473–4479). Piscataway, USA: IEEE Press.

Ye, F.Y., & Rousseau, R. (2008). The power law model and total career *h*-index sequences. Journal of Informetrics, 2(4), 288–297.

Yehezkel, R., & Lerner, B. (2009). Bayesian network structure learning by recursive autonomy identification. Journal of Machine Learning Research, 10(Jul), 1527–1570.

Appendix

In this appendix, we provide a small, practical example on how to perform the calculations of GBNs and GAs. Given a set of predictive variables $\boldsymbol{X} = \{A, B\}$ and response variables $\boldsymbol{Y} = \{C, D\}$, we learn a GBN using a GA. This GA explores different GBN structures and selects the best GBN structure that minimizes the distance between real and predicted response variable values. All steps required to achieve the above goal are detailed here.

GA—Initial Population

The first step of a GA consists of the initialization of possible solutions. The initial population of our GA (20 individuals) is randomly generated from a Bernoulli distribution. Each individual of the population represents a GBN structure described by an adjacency matrix. The entries a_{ij} of the adjacency matrix represent arcs among nodes, such that $a_{ij} = 1$ if, and only if, an arc exists from node i to node j, and $a_{ij} = 0$ otherwise. Here, we show some examples of adjacency matrices.

$$\begin{array}{cc} \textit{Individual-1} & \textit{Individual-2} \\ \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

$$\begin{array}{ccc} \textit{Individual-3} & \ldots & \textit{Individual-20} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} & \ldots & \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

DOI: 10.1002/asi

Once the population is generated, the following GA steps improve initial individuals through repetitive application of the selection, crossover, and mutation operators.

GA—Fitness Function

The next step is to compute a fitness score (in our case, the Mahalanobis distance between real and predicted responses) for each individual of the population using Equation (9). The complexity of this function lies in the calculation of the predicted values because it is necessary to learn the parameters of a GBN given a fixed structure and then forecast the predicted values given specific evidences. For the sake of simplicity, we only show how to compute the fitness score associated with *Individual-1*.

Learning the GBN Parameters The GBN is defined as a pair of elements: the structure represented by the selected adjacency matrix (see Figure 4) and its joint probability distribution, which can be specified by the product of a set of conditional probability distributions, that is,

$$f(a,b,c,d) = f(a)f(b)f(c|a)f(d|a,b)$$

where

$$f(a) \sim \mathcal{N}(\mu_A, v_A) \quad f(c|a) \sim \mathcal{N}(\mu_C + \beta_{AC}(a-\mu_A), v_C)$$
$$f(b) \sim \mathcal{N}(\mu_B, v_B) \quad f(d|a,b) \sim \mathcal{N}(\mu_D + \beta_{AD}(a-\mu_A) + \beta_{BD}(b-\mu_B), v_D)$$

The parameters involved in this representation are $\boldsymbol{\mu}^T = (\mu_A, \mu_B, \mu_C, \mu_D)$, $\boldsymbol{v}^T = (v_A, v_B, v_C, v_D)$ and $\boldsymbol{\beta}^T = (\beta_{AC}, \beta_{AD}, \beta_{BD})$. These values can be computed from the training data. The first vector represents the mean values of variables, the second vector includes the conditional variance of a variable given its parents (they can be calculated using Equation [5]), and the third vector represents the regression coefficients between variables. For $\mu_A = \mu_B = \mu_C = \mu_D = 0.00$, $v_A = v_B = v_C = v_D = 1.00$, $\beta_{AC} = 1.00$, $\beta_{AD} = 0.20$, and $\beta_{BD} = 0.80$, we get the following conditional probability distributions:

$$f(a) \sim \mathcal{N}(0.00, 1.00) \quad f(c|a) \sim \mathcal{N}(a, 1.00)$$
$$f(b) \sim \mathcal{N}(0.00, 1.00) \quad f(d|a,b) \sim \mathcal{N}(0.20a + 0.80b, 1.00)$$

The GBN model can be also defined by the mean vector and the covariance matrix $\boldsymbol{\Sigma}$. In order to calculate the covariance matrix, we first calculate its inverse matrix, the $\boldsymbol{W}$ precision matrix, using Equation (6). After four iterations, we get

$$\boldsymbol{W} = \begin{pmatrix} \frac{1}{v_A} + \frac{\beta_{AC}^2}{v_C} + \frac{\beta_{AD}^2}{v_D} & \frac{\beta_{AD}\beta_{BD}}{v_D} & \frac{-\beta_{AC}}{v_C} & \frac{-\beta_{AD}}{v_D} \\ \frac{\beta_{BD}\beta_{AD}}{v_D} & \frac{1}{v_B} + \frac{\beta_{BD}^2}{v_D} & 0 & \frac{-\beta_{BD}}{v_D} \\ \frac{-\beta_{AC}}{v_C} & 0 & \frac{1}{v_C} & 0 \\ \frac{-\beta_{AD}}{v_D} & \frac{-\beta_{BD}}{v_D} & 0 & \frac{1}{v_D} \end{pmatrix}$$

Finally, with the number above, we have

$$\boldsymbol{W} = \begin{pmatrix} 2.04 & 0.16 & -1.00 & -0.20 \\ 0.16 & 1.64 & 0.00 & -0.80 \\ -1.00 & 0.00 & 1.00 & 0.00 \\ -0.20 & -0.80 & 0.00 & 1.00 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.00 & 0.00 & 1.00 & 0.20 \\ 0.00 & 1.00 & 0.00 & 0.80 \\ 1.00 & 0.00 & 2.00 & 0.20 \\ 0.20 & 0.80 & 0.20 & 1.68 \end{pmatrix}$$

Exact Evidence Propagation in GBN

Once the GBN is induced, we predict the response variables values using Equations (8) and (9). These equations require the submatrices of the covariance matrix $\boldsymbol{\Sigma}$, the mean vector of predictive variables $\boldsymbol{\mu}_X$, the mean vector of response variables $\boldsymbol{\mu}_Y$, and the evidences $\boldsymbol{x}$ associated with the predictive variables. In the example, given the evidence $\boldsymbol{x}^T = (A = 1.00, B = 3.00)$, the response variables $\boldsymbol{y}^T = (C, D)$ are calculated as follows:

$$\begin{aligned} \boldsymbol{\mu}^{Y|X=x} &= \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_X) \\ &= \begin{pmatrix} 0.00 \\ 0.00 \end{pmatrix} + \begin{pmatrix} 1.00 & 0.00 \\ 0.20 & 0.80 \end{pmatrix} \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix}^{-1} \\ &\quad \left(\begin{pmatrix} 1.00 \\ 3.00 \end{pmatrix} - \begin{pmatrix} 0.00 \\ 0.00 \end{pmatrix} \right) \\ &= \begin{pmatrix} 1.00 \\ 2.60 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \boldsymbol{\Sigma}^{Y|X=x} &= \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} \\ &= \begin{pmatrix} 2.00 & 0.20 \\ 0.20 & 1.68 \end{pmatrix} - \begin{pmatrix} 1.00 & 0.00 \\ 0.20 & 0.80 \end{pmatrix} \\ &\quad \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix}^{-1} \begin{pmatrix} 1.00 & 0.20 \\ 0.00 & 0.80 \end{pmatrix} \\ &= \begin{pmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{pmatrix} \end{aligned}$$

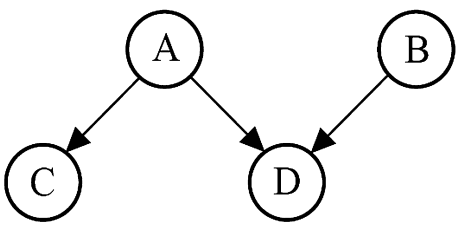

FIG. 4. GBN structure of *Individual-1*.

DOI: 10.1002/asi

After computing these equations, we conclude that variable C follows a Gaussian distribution $\mathcal{N}(1.00, 1.00)$, whereas variable D follows a distribution $\mathcal{N}(2.60, 1.00)$, given the evidences $A = 1.00$ and $B = 3.00$. Thus, the vector of predicted values is $\boldsymbol{y'^T} = (C = 1.00,\ D = 2.60)$.

Mahalanobis Distance as Fitness Score

The Mahalanobis distance between real and predicted response variable values is used as the fitness score for each individual of the search population in the GA. This function (see Equation [9]) requires three parameters: the covariance matrix of the response variables and two vectors representing the real and predicted values of the response variables. Given the vector of actual values for the response variables, $\boldsymbol{y}^T = \{C = 1.00,\ D = 2.50\}$, the vector of predicted values $\boldsymbol{y'^T} = \{C = 1.00,\ D = 2.60\}$, and the covariance matrix of the response variables, the Mahalanobis distance (MD) can be calculated as

$$\begin{aligned} MD(\boldsymbol{y}, \boldsymbol{y'}) &= \sqrt{(\boldsymbol{y} - \boldsymbol{y'})^T \Sigma_{YY}^{-1} (\boldsymbol{y} - \boldsymbol{y'})} \\ &= \sqrt{\left(\begin{pmatrix}1.00\\2.50\end{pmatrix} - \begin{pmatrix}1.00\\2.60\end{pmatrix}\right)^T \begin{pmatrix}2.00 & 0.20\\0.20 & 1.68\end{pmatrix}^{-1} \left(\begin{pmatrix}1.00\\2.50\end{pmatrix} - \begin{pmatrix}1.00\\2.60\end{pmatrix}\right)} \\ &= 0.07761505 \end{aligned}$$

The value just shown represents the Mahalanobis distance related to one case of the testing set. The final fitness score of *Individual-1* corresponds to the mean value of all Mahalanobis distances associated with the testing cases.

The processes of inducing the GBN model, propagating new evidences, and calculating the Mahalanobis distance is repeated for all individuals within the population, achieving a fitness score for each of them.

GA—Selection Criterion

After computing the fitness scores of the current population, we proceed to select which individuals will mate to create new individuals. We use an elitist strategy, which chooses the best $N/2$ individuals from the population for reproduction. Considering that the aim is to minimize the distance between real and predicted response variable values, the lower the fitness score, the fitter an individual is. Given the fitness scores of the current population, the selected individuals are:

Individual	Fitness	Selected	Individual	Fitness	Selected
Individual-1	0.12	Yes	*Individual-2*	0.85	No
Individual-4	0.51	Yes	*Individual-3*	0.72	No
Individual-7	0.48	Yes	*Individual-5*	0.61	No
Individual-8	0.22	Yes	*Individual-6*	0.82	No
Individual-11	0.09	Yes	*Individual-9*	1.32	No
Individual-12	0.21	Yes	*Individual-10*	0.97	No
Individual-15	0.43	Yes	*Individual-13*	0.74	No
Individual-18	0.26	Yes	*Individual-14*	0.63	No
Individual-19	0.43	Yes	*Individual-16*	0.88	No
Individual-20	0.11	Yes	*Individual-17*	1.05	No

Finally, the 10 selected individuals (parents) are moved into a mating pool where they are combined by crossover and mutation operations.

GA—Crossover and Mutation

The selected parents are randomly mated in pairs to create 10 new children by combining their genotypic information. Given the codification of the parents as strings, we randomly choose, with a fixed probability, a crossover point at which the information is exchanged. This is called single-point crossover. 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. Here, we show how the crossover operator in the pair formed by *Individual-1* and *Individual-20* works.

Individual-1

$$\begin{pmatrix}0&0&1&1\\0&0&0&1\\0&0&0&0\\0&0&0&0\end{pmatrix}$$

$(a_{11}, a_{12}, a_{13}, a_{14}, a_{21}, \ldots, a_{42}, a_{43}, a_{44})$

$(0, 0, 1, 1, 0, 0, 0, 1, 0, 0, \mathbf{0}, 0, 0, 0, 0, 0)$

Individual-20

$$\begin{pmatrix}0&1&0&1\\0&\mathbf{0}&1&0\\0&0&0&1\\0&0&0&0\end{pmatrix}$$

$(b_{11}, b_{12}, b_{13}, b_{14}, b_{21}, \ldots, b_{42}, b_{43}, b_{44})$

$(0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0)$

Assuming a crossover point in the eighth bit, new offspring are

$$(0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0)$$
$$(0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, \mathbf{0}, 0, 0, 0)$$

Finally, the mutation operator introduces some extra variability into the new offspring by random changes on a low probability basis (mutation probability). A bit from the offspring binary strings is randomly chosen and flipped in the example. That is,

First offspring

$(0, 0, 1, 1, 0, 0, \mathbf{0}, 1, 0, 0, 0, 1, 0, \mathbf{0}, 0, 0)$

Second offspring

$(0, \mathbf{1}, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$

First mutated offspring

$(0, 0, 1, 1, 0, \mathbf{0}, \mathbf{1}, 1, 0, 0, 0, 1, 0, 0, 0, 0)$

Second mutated offspring

$(0, \mathbf{0}, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$

GA—Merging Procedure

Once the new individuals are generated, their fitness score is also calculated. After that, an elitist strategy yields the new population of individuals by combining the 20 individuals from the previous population with the 10 new ones. Only the top-ranked 20 individuals are moved into the new population.

GA—Stopping Criteria

The GA finishes when a set of conditions are fulfilled. Here, reaching 20 generations or no improvement over five sequential generations constitutes our stopping criteria.

DOI: 10.1002/asi

The individual with the lowest fitness score in the final population is considered to be the solution to the optimization problem, that is, it is the GBN that minimizes the distance between real and predicted response variable values using $\{A, B\}$ as predictive variables and $\{C, D\}$ as response variables.

The same process is run using all different combinations of predictive and response variables. After reaching all different solutions, the result is a set of three optimal GBNs. Each model is associated with a different combination of predictive and response variables, that is, one predictive variable and three response variables, two predictive variables and two response variables, and, finally, three predictive variables and one response variable. From these optimal solutions, only the best GBN model is retained. The BIC value is used to quantitatively assess which of the models to retain. The resulting model is expected to uncover the best core set of relevant indices with the highest predictive power in forecasting bibliometric indices.

DOI: 10.1002/asi