Knowledge-Based Systems 50 (2013) 87–99

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Ranking evaluation of institutions based on a Bayesian network having a latent variable

Jun-Seong Kim, Chi-Hyuck Jun *

Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang 790-784, Republic of Korea

ARTICLE INFO

Article history:
Received 5 April 2012
Received in revised form 29 April 2013
Accepted 22 May 2013
Available online 31 May 2013

Keywords:
Ranking estimation
Linear Gaussian model
Structure learning
Gibbs sampling
Multiple search
Causal discovery

ABSTRACT

This paper proposes a new probabilistic graphical model which contains an unobservable latent variable that affects all other observable variables, and the proposed model is applied to ranking evaluation of institutions using a set of performance indicators. Linear Gaussian models are used to express the causal relationship among variables. The proposed iterative method uses a combined causal discovery algorithm of score-based and constraint-based methods to find the network structure, while Gibbs sampling and regression analysis are conducted to estimate the parameters. The latent variable representing ranking scores of institutions is estimated, and the rankings are determined by comparing the estimated scores. The interval estimate of the ranking of an institution is finally obtained from a repetitive procedure. The proposed procedure was applied to a real data set as well as artificial data sets.

1. Introduction

Ranking higher education institutions began in the United States in the early 20th century, and it has become increasingly more influential with time. Most ranking systems select the performance indicators (research, education, etc.), assign a weight to each indicator and calculate the weighted score for each institution; this is the 'weight-and-sum' approach. However, this method sometimes becomes controversial because assigning weights to performance indicators is subjective. Thus demand for a new quantitative method of ranking is on the rise although there are many issues other than raking methods as pointed out in Vught and Westerheijden [34].

Some studies have used approaches other than the weight-and-sum approach to rank institutions. These methods include an empirical Bayes approach to ranking schools based on student achievement [18], Bayesian analysis for ranking institutions [15], a latent-variable technique for university ranking based on several indicators [16] and the analytic hierarchy process for determining weights of indicators [23]. In addition, a method based on a supervised naïve Bayes structure uses the mixture of truncated exponentials, which applies the rank information of an expert [12]. A Bayesian latent variable model has been proposed for estimating the top ranked SNPs detected from genetic association studies [13], which may also be applied to the institution ranking problem.

When using the ranking evaluation systems other than the weight-and-sum approach, a structure analysis is necessary which considers the relationship among performance indicators. However, few studies have been conducted on the structure analysis for ranking evaluation. The ordinary modeling techniques for the structure analysis are structural equation models (SEMs) based either on linear structure relationships or partial least squares and Bayesian networks (BNs) [20]. Especially, BNs can be used as probabilistic inference engines, building models of domains that have intrinsic uncertainty. BNs are graphical models based on the notion of conditional independence that subsumes a wide range of statistical models including regression models, factor analysis models, and structural equation models. In a BN, a directed acyclic graph (DAG) represents a set of conditional independence constraints among a given number of variables and their related conditional probability distributions. The procedures for developing BNs involve learning the structure (the relationships between variables) first, and then parameterizing the associated conditional distributions. Graphical models, in particular those based on DAGs, have natural causal interpretations and thus form a language in which causal concepts can be discussed and analyzed in precise terms [21]. The conditional independence assumptions in a BN yield models more compact than those based on full joint probability distributions, thus reducing computational complexity when the number of variables is large [30]. Lately, BNs have been used in various applications such as risk management [3], resource allocation decisions [10,11], IT implementation [20], species distribution [1], higher education [12], and health risk assessment [22].

* Corresponding author. Address: Industrial and Management Engineering, POSTECH, Republic of Korea. Tel.: +82 542792197.
E-mail addresses: soulloan@postech.ac.kr (J.-S. Kim), chjun@postech.ac.kr (C.-H. Jun).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.05.010

In this paper, we propose a new Bayesian network model which has an unobservable latent variable that affects all other observable variables. To solve the ranking problem, the latent variable represents the ranking variable to be finally estimated, and the observable variables indicate performance indicators of each institution. Our first task is to identify the causal relation among these variables, and the second task is to determine the institution rankings in terms of intervals.

Causal discovery algorithms (CDAs) for a BN can be classified into three different categories: the score-based approach, the constraint-based approach, and the combined approach [35]. Score-based methods select a model with the highest posterior probability when the prior of each model is given. Constraint-based methods use statistical methods to detect associations and independencies among variables. Each method has a trade-off between time complexity and accuracy. Score-based methods generally provide accuracy but their time complexity is very high. Constraint-based methods give much lower time complexity, but they may not be accurate if conditional independence (CI) tests fail. The combined methods use both concepts to build the graph by compromising the benefit in terms of time complexity and accuracy. These methods cannot be directly applied to our new graphical model which has a latent variable, so a new approach needs to be developed for our purpose.

The proposed approach consists of two phases and each phase is repeated until the network structure converges. First, latent rankings are obtained using Gibbs sampling and gradient descent from a given network structure. Second, the updated network structure is found by the revised version of Multiple Search (MS) algorithm [7] using the latent rankings. The final rankings in terms of interval estimates are obtained based on the convergent structure.

The rest of the paper is organized as follows. The BN under consideration is modeled in Section 2. The proposed method of learning the BN is described in Section 3. The proposed method is applied to real data in Section 4. Section 5 reports on a further experiment with artificial data to check the accuracy and consistency of the proposed new Bayesian network model. Conclusions are presented in Section 6.

2. Bayesian network having a latent variable

Our ranking problem is to assign a ranking to each of the given number of institutions based only on their performance indicators. For this purpose we will consider a new type of BN as shown in Fig. 1. The difference from the usual BN is to contain an unobservable latent variable that affects all other observable variables. Let X_i ($i = 1,2,\ldots,m$) be the ith variable where m is the number of variables. These variables are performance indicators for our ranking problem. They are assumed to be observable for each institution, and they may have causal relationships among themselves. Let Z be an unobservable latent variable that affects all other variables. Without the latent variable Z, the network will be the same as a usual graphical model. For our ranking problem, Z_j ($j = 1,2,\ldots,n$) represents the unobservable ranking score of the jth institution, where n is the number of institutions. Hence, our problem is to estimate the ranking scores of all institutions by considering the causal relationship among observable performance indicators. Once the ranking scores of all institutions are estimated, the ranking will be determined by comparing the scores' magnitudes.

2.1. Linear Gaussian model

Consider an arbitrary DAG in which node i represents a continuous random variable X_i that has a Gaussian distribution. Then, under the linear Gaussian model, the conditional distribution of X_i given its parent node pa_i is also Gaussian:

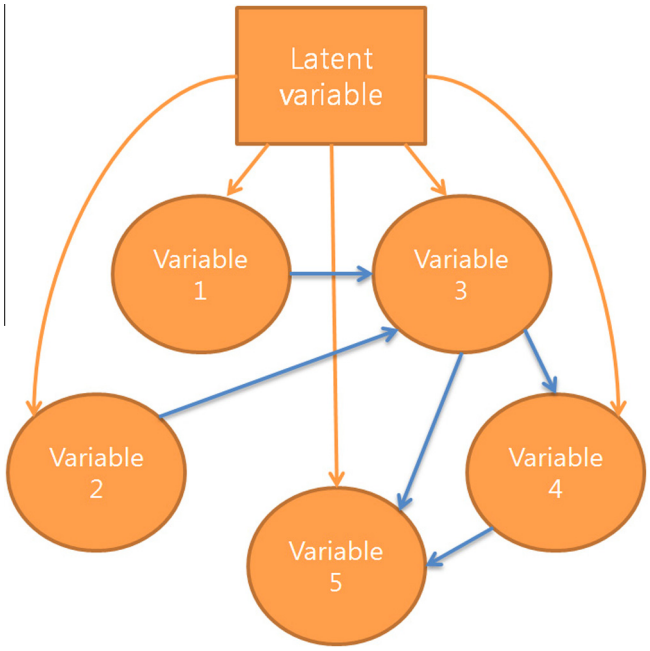

Fig. 1. An example of our graphical model having a latent variable.

$$X_i|X_{pa_i} \sim N\left(\sum_{j\in pa_i} w_{ij}X_j + b_i,\ v_i\right) \tag{1}$$

where w_{ij} and b_i are parameters that govern the relationship of $X_{pa_i} \to X_i$, and v_i is the conditional variance of X_i.

A DAG 'G' depicts a set of variables, $X_1,X_2,\ldots,X_m$, and their relationships. Each node of G is a variable, and the directed arcs represent the parental relationships among the variables. The joint probability distribution of G is

$$p(x_1,x_2,\ldots,x_m|G) = \prod_{i=1}^{m} p(x_i|x_{pa_i},G) \tag{2}$$

where $p(x_1,x_2,\ldots,x_m|G)$ represents the probability distribution of a specific combination $x_1,x_2,\ldots,x_m$ from the variables $X_1,X_2,\ldots,X_m$, and x_{pa_i} is a vector that represents the list of direct parents of X_i, as depicted by G. BNs are locally structured, meaning that each node interacts only with its parent nodes.

For a given structure G under consideration, we estimate its parameters in Eq. (1) and derive the distribution of the latent variable Z_j. Because our graphical model includes a latent variable, the existing method cannot be applied. So, we developed a new procedure, which will be described in Section 3.

2.2. Learning Bayesian networks from data

If the structure of the BN is not known, the underlying structure of the BN given by G must be learned. This structure includes the specifications for the conditional independence assumptions among the variables of the model and the parameters. Many DAGs may determine the same joint probability distribution. Therefore, the family of all DAGs with a given set of vertices is naturally partitioned into Markov-equivalence class, each class being associated with a unique statistical model. This means that the structure of