Available online at www.sciencedirect.com
SCIENCE DIRECT®
Decision Support Systems 38 (2004) 451–472
Decision Support Systems
www.elsevier.com/locate/dsw

# Data mining of Bayesian networks using cooperative coevolution

Man Leung Wong[a,*], Shing Yan Lee[b], Kwong Sak Leung[b]

[a] *Department of Information Systems, Lingnan University, Tuen Mun, Hong Kong, China*
[b] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China*

Received 1 April 2003; accepted 23 July 2003
Available online 19 September 2003

**Abstract**

This paper describes a novel data mining algorithm that employs cooperative coevolution and a hybrid approach to discover Bayesian networks from data. A Bayesian network is a graphical knowledge representation tool. However, learning Bayesian networks from data is a difficult problem. There are two different approaches to the network learning problem. The first one uses dependency analysis, while the second approach searches good network structures according to a metric. Unfortunately, the two approaches both have their own drawbacks. Thus, we propose a novel algorithm that combines the characteristics of these approaches to improve learning effectiveness and efficiency. The new learning algorithm consists of the conditional independence (CI) test and the search phases. In the CI test phase, dependency analysis is conducted to reduce the size of the search space. In the search phase, good Bayesian networks are generated by a cooperative coevolution genetic algorithm (GA). We conduct a number of experiments and compare the new algorithm with our previous algorithm, Minimum Description Length and Evolutionary Programming (MDLEP), which uses evolutionary programming (EP) for network learning. The results illustrate that the new algorithm has better performance. We apply the algorithm to a large real-world data set and compare the performance of the discovered Bayesian networks with that of the back-propagation neural networks and the logistic regression models. This study illustrates that the algorithm is a promising alternative to other data mining algorithms.


*Keywords:* Cooperative coevolution; Evolutionary computation; Data mining; Bayesian networks

## 1. Introduction

Data mining is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [19]. The whole process of data mining consists of several steps. Firstly, the problem domain is analyzed to determine the objectives. Secondly, data is collected and an initial exploration is conducted to understand and verify the quality of the data. Thirdly, data preparation such as selection is made to extract relevant data sets from the database. The data is preprocessed to remove noise and to handle missing data values. Transformation may be performed to reduce the number of variables under consideration. A suitable data mining algorithm is then employed on the prepared data to discover knowledge represented in different representations such as decision trees, rules, and Bayesian networks. Finally, the result of data mining is interpreted and evaluated. If the

* Corresponding author.
*E-mail addresses:* mlwong@ln.edu.hk (M.L. Wong), sylee@cse.cuhk.edu.hk (S.Y. Lee), ksleung@cse.cuhk.edu.hk (K.S. Leung).

0167-9236/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-9236(03)00115-5

discovered knowledge is not satisfactory, these steps will be iterated. The discovered knowledge is then applied in decision making.

Recently, there is increasing interest in discovering knowledge represented in Bayesian networks [16,25,47,54,55] because Bayesian networks can handle incomplete data sets and facilitate the combination of domain knowledge and data. Moreover, Bayesian networks provide an efficient way for avoiding the over fitting problem and allow one to learn about causal relationships [25]. In this paper, we propose a novel data mining algorithm that employs cooperative coevolution and a hybrid approach to learn knowledge represented in Bayesian networks from data. A Bayesian network is a graphical representation that depicts conditional independence (CI) among random variables in the domain and encodes the joint probability distribution. A Bayesian network is composed of a structure and a number of conditional probabilities as shown in Fig. 1. With a network at hand, probabilistic inference can be performed to predict the outcome of some variables based on the observations of others.[1] In light of this, Bayesian networks are widely used in diagnostic and classification systems. For example, they are used for diagnosing diseases in muscles, nerve, and lymph nodes [28,39]. Besides, they are also used in information retrieval [26] and printer troubleshooting problems [27].

The main task of learning Bayesian networks from data is to automatically find directed edges between the nodes. Once the network structure is constructed, the conditional probabilities are readily calculated based on the data. In the literature, there are two main approaches to learning network structure from data [14]. The first one is the dependency analysis approach [14,50]. Since a Bayesian network describes conditional independence, we could make use of dependency test results to construct a Bayesian network structure that conforms to our findings. The second one, called the score-and-search approach [24,30,34], uses a metric to evaluate a candidate network structure. With the metric, a search algorithm is employed to find a network structure, which has the best score. Thus, the learning problem becomes a search problem. Unfortunately, the two approaches both have their own drawbacks. For the former approach, an exponential number of dependency tests should be performed. Moreover, some test results may be inaccurate [50]. For the latter approach, since the search space is huge, some Bayesian network structure learning algorithms [30] apply greedy search heuristics, which may easily make the algorithms get stuck in a local optimum [24].

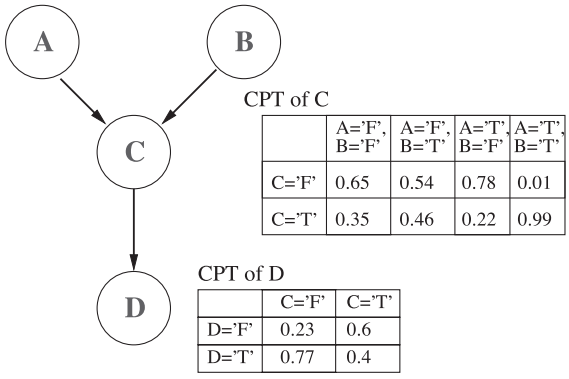


Fig. 1. A Bayesian network example.

In this work, a hybrid approach is developed for the network structure learning problem. Simply put, dependency analysis results are used to reduce the search space of the score-and-search process. With such reduction, the search process would take less time for finding the optimal solution. A modular decomposition evolutionary search approach, cooperative coevolution [42–44], is employed to search for good Bayesian network structures. Our new algorithm for learning Bayesian network structures is called Cooperative Coevolution Genetic Algorithm (CCGA). We have conducted a number of experiments and compared CCGA with our previous algorithm, Minimum Description Length and Evolutionary Programming (MDLEP). The empirical results illustrate that CCGA outperforms MDLEP. Moreover, it is found that CCGA executes much faster than MDLEP, which is very important for real-world applications. We evaluate CCGA on a real-world data set of direct marketing and compare the performance of the discovered Bayesian networks with that of the back-propagation neural networks and the logistic regression models.

[1] There are several commercial and free software such as HUGIN [32], Bayesian Network tools in Java [10] and Microsoft Belief Network Tools [37] that can perform probabilistic reasoning given a Bayesian network.