Expert Systems with Applications 37 (2010) 5470–5475

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Improving algorithms for structure learning in Bayesian Networks using a new implicit score

Lobna Bouchaala [a], Afif Masmoudi [b], Faiez Gargouri [c], Ahmed Rebai [a,*]

[a] Bioinformatics Unit, Centre de Biotechnologie de Sfax, P.O. Box 1177, 3018 Sfax, Tunisia
[b] Faculty of Science of Sfax, Department of Mathematics, Sfax, Tunisia
[c] Higher Institute of Computer and Multimedia of Sfax, Sfax, Tunisia

ARTICLE INFO

Keywords:
Bayesian Network
Implicit method
Implicit score
Structure-learning algorithm
Modeling
Breast cancer

ABSTRACT

Learning Bayesian Network structure from database is an NP-hard problem and still one of the most exciting challenges in machine learning. Most of the widely used heuristics search for the (locally) optimal graphs by defining a score metric and employs a search strategy to identify the network structure having the maximum score. In this work, we propose a new score (named implicit score) based on the Implicit inference framework that we proposed earlier. We then implemented this score within the K2 and MWST algorithms for network structure learning. Performance of the new score metric was evaluated on a benchmark database (ASIA Network) and a biomedical database of breast cancer in comparison with traditional score metrics BIC and BD Mutual Information. We show that implicit score yields improved performance over other scores when used with the MWST algorithm and have similar performance when implemented within K2 algorithm.

1. Introduction

Bayesian Network (BN) is one of the most popular models in data mining technologies. A BN is a Directed Acyclic Graph (DAG), where the nodes are random variables and where the arcs specify the independence assumptions between these variables. During the past few years, several algorithms have been developed for learning the structure of BN from a database (Neapolitan, 2003). An important class of such algorithms is known as the score metric-based methods. These methods first define quantity metric that measures the quality of the network structures or the conditional dependence among variables given the database; second, they apply a search procedure to identify a structure that maximizes the metric over the structure space. Many scores metrics have been proposed and employed with a heuristic search strategy to identify the network with the maximum score. The most used algorithm, named the K2 algorithm, was proposed by cooper and Hersovits (1992) with a Bayesian Dirichlet (BD) score metric. Other authors used this algorithm with other score metrics such as the likelihood-equivalence Bayesian Dirichlet metric (BDe) (Heckerman, Geiger, & Chickering, 1994) and the minimum description length (MDL) metric (Bouckaert, 1993; Lam & Bacchus, 1993; Rissanen, 1978) showing an improvement in the performance. Recently, Yun and Keong (2004) proposed another formulation of the MDL score, which they called IMDL (Improved MDL) .The BIC score metric proposed by Akaike (1973, 1979) has been also used and implemented in several computer packages. Recently, Chen (2008) developed a new structure-learning algorithm based on scoring search methods and information theory.

These score metrics have been widely compared and used. Shulin and Chang (2002) compared the performance of five score metrics using the K2 algorithm. Acid and de Campos (2001) introduced a new score and assessed its performance within the K2 algorithm in comparison with other scores.

In this work, we propose a new score metric that we named implicit score (IS). This score is based on the Implicit inference framework proposed by Hassairi, Masmoudi, and Kokonendji (2005) and applied to probability learning in Bayesian Networks by BenHassen, Masmoudi, and Rebai (2008). We implemented this score within the K2 and MWST algorithms and compared its performance to the traditional score metrics (BD, BIC and Mutual Information) available in most implementation.

This paper is organized as follows: In Section 2, we introduce BNs and relevant concepts. In Section 3, we describe the Implicit method and the new proposed score. Section 4 presents the experimental results. Finally, Section 5 presents our conclusions.

2. Background

2.1. Bayesian Network definition

A Bayesian Network is a graphical model for probabilistic relationships among a set of variables $X = \{X_1, \ldots, X_n\}$. If X_i is a direct

* Corresponding author. Tel.: +216 74871816; fax: +216 74875818.
E-mail address: ahmed.rebai@cbs.rnrt.tn (A. Rebai).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.02.065

'cause' of X_j, then we draw an edge from X_i to X_j. The set E of vertices (i.e. edges) represents all the dependencies links between the variables. So the network (X, E) is a Directed Acyclic Graph (DAG).

A Bayesian Network satisfies the Markov condition, which states that the value of a variable is conditionally independent of its non-descendents given its parents.

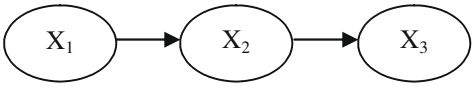
For example, consider a set of three variables X_1, X_2, X_3, the joint probability distribution of (X_1, X_2, X_3) is given by

$$P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)$$

If X_3 is independent of X_2 given X_3, then

$$P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_2)$$

And



In a Bayesian Network, the DAG is called the structure, and the values of the conditional probability distributions are called the parameters and denoted θ_{ijk}; θ_{ijk} is the conditional probability of X_i in the state k given that its parents $\mathrm{Pa}(X_i)$ are in the state j.

$$\theta_{ijk} = P(Xi = k|\mathrm{Pa}(Xi) = j) \quad (1)$$

The learning task of the BNs can be separated into two subtasks, structure learning, that is to identify the topology of the network, and parameters learning, for calculating the numerical parameters (conditional probabilities) for a given network topology. Learning BN structure from the data has been proven to be an NP-hard problem; Robinson (1977) showed that the number $r(n)$ of possible structures for Bayesian Network having n nodes is given by the recurrence formula:

$$r(n) = \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{i!(n-i)!} 2^{i(n-1)} r(n-i) \quad (2)$$

Searching for the best structure is difficult because the search space increases exponentially with the number of variables. Most algorithms used for structure learning are designed to reduce the search space.

One of the most used algorithms is the K2 algorithm (Cooper & Hersovits, 1992). The K2 algorithm proceeds as follows: we assume an initial ordering of the nodes such that, if X_j proceeds X_i in the order, an arc from X_j to X_i is not allowed. This initial ordering can be based on prior expert knowledge or on some data (ordering by variance for discretized quantitative variables). Let $Pred(X_i)$ be the set of nodes that precede X_i in the ordering, we initially set the parents $\mathrm{Pa}(X_i)$ of X_i to empty and compute the score $B(D, X_i, \mathrm{Pa}(X_i))$. Next, we visit the nodes in sequence according to the ordering. When we visit X_i, we determine the node in $Pred(X_i)$, which most increases score $B(D, X_i, \mathrm{Pa}(X_i))$. We "greedily" add this node to $\mathrm{Pa}(X_i)$. We continue doing this until the addition of no node increases the score $B(D, X_i, \mathrm{Pa}(X_i))$ or the maximum number of parents for that particular node has been reached.

Algorithm 1.

The K2 algorithm.

```
Input:  – A set of N nodes X = {X1, ..., XN}
        – An ordering on the nodes
        – A data set D
        – M the maximum in degree of a node
Output: For each node, a printout of the parent of the node
For i = 1 to N do
  πi = Ø
  ScoreOld = score (D, Xi, πi)
  OkToProceed = True
  While OkToProceed && |πi| < M do
    Let Xj be the node ind Pred(Xi) − πi that maximizes ((D,
  Xi, πi) union Xj)
    ScoreNew = score((D, Xi, πi) union Xj)
    If ScoreNew > ScoreOld
    Then
      ScoreOld = ScoreNew
      πi = πi union Xj
    Else
      OkToProceed = false
    end If
  end While
end For
```

Chow and Liu (1968) proposed a method derived from the Maximum Weight Spanning Tree (MWST). This method associates a weight to each potential edge $X_i - X_j$ of the tree. This weight may be the Mutual Information between the variables X_i and X_j as proposed by Chow and Liu (1968), or the local variation of the score when choosing X_j as a parent of X_i. Given the weight matrix, one can use standard algorithms as Kruskal algorithm (Kruskal, 1956) to solve the problem of the spanning tree maximum weight and obtain a directed tree by choosing a root and then browsing the tree by an in-depth research.

Algorithm 2.

The MWST algorithm.

```
Input:  – A set of N nodes X = {X1, ..., XN}
        – A data set D
Output: For each node, a printout of the parent of the node
Repeat
1. Calculate the P(xi, xj) for all pair in the boundary from the
   overall distribution
2. Calculate the branch weight I(Xi, Xj) for all of the edges in
   the boundary
3. Assign the edges with the largest I to the tree if it is not
   forming a cycle
Until n − 1 edges have been selected
```

2.2. Score metrics

In what follows, we consider a DAG with n nodes. For each node i, we associate a random variable X_i having r_i states. So the joint distribution of $(X_1, \ldots, X_n)$ is a multinomial distribution.

Given a structure B, we denote by N_{ijk} the number of cases in D where X_i is in its jth state and its parents are in their kth state, by $N_{ij} = \sum_{k=1}^{r} N_{ijk}$ the number of cases in D where X_i is in its jth state, r_i the number of states of variable X_i, q_i the number of parent configurations of X_i and $\theta_{ijk} = P(X_i = k|\mathrm{Pa}(X_i) = j)$ for $k = 1, \ldots, r$ and $\sum_{k=1}^{r} \theta_{ijk} = 1$.

A scoring criterion for a DAG is a function that assigns a value to each DAG under consideration based on the data. Cooper and Hersovits (1992) proposed a score based on a Bayesian approach (Bayesian with Dirichlet priori called BD). Starting from a prior distribution on the possible structures $P(B)$, the objective is to express the posterior probability of all possible structures ($P(B|D)$ or simply $P(B, D)$) conditional on a dataset D:

$$S_{BD}(B, D) = P(B, D) = \int_{\Theta} P(D|\Theta, B)P(\Theta|B)P(B)d\Theta$$
$$= P(B) \int_{\Theta} P(D|\Theta, B)P(\Theta|B)d\Theta \quad (3)$$