An Improved Bayesian Networks Learning Algorithm Based on Independence Test and MDL Scoring

Junzhong Ji, Jing Yan, Chunnian Liu
College of Computer Science and Technology, Beijing University of Technology
Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology
Beijing 100022, China
jjz01@bjut.edu.cn

Ning Zhong
Department of Information Engineering, Maebashi Institute of Technology,
460-1 Kamisadori-Cho, Maebashi-City, 371-0816, Japan
zhong@maebashi-it.ac.jp

Abstract

In recent years, more and more people studied the Bayesian networks learning algorithm that integrates independence test with scoring metric. Based on the proposed hybrid algorithm I-B&B-MDL, a modified method is developed. There are two major contributions. Firstly, order-0 and partial order-1 independence tests are used to obtain an original graph of the network, which reduces the number of independence tests and database passes while effectively restricting the search space. Secondly, by means of the heuristic knowledge of mutual information, sort order for candidate parent nodes increases the cut-offs of the B&B search tree and accelerates search process. The experimental results show that the modified algorithm has high accuracy, and is more efficient in time complexity than other algorithms.

1 Introduction

A Bayesian network (BN) is an annotated directed acyclic graph that encodes a joint probability distribution over a finite set of random variables. In uncertainty knowledge field, BN is a powerful knowledge representation and reasoning tool. Compared with other information models, BN has several advantages, as follows.

- BN represents clearly the probability distribution relationships for the domain variables in a graph way, and asserts the conditional independence between variables or variable sets. Thus, it is easy to unite the hypotheses and the observed data during the reasoning.
- As BN is a kind of probabilistic semantic knowledge representation, it is easy to understand the field knowledge on the whole. It is also convenient for explaining the meaning of local area, such as, node variables, directed arcs, graph branches and so on.
- A BN's structure uniforms qualitative and quantitative knowledge representation. If considering a directed arc as a qualitative depiction of dependence, a conditional probability then reflects intensity of this relationship.

As mentioned above, BN can depict and discover much convincing the probability dependence. Thus, it has been widely used in many fields, such as, data mining, pattern recognition, intelligent tutoring system, medical diagnoses and so on. This paper focuses on BN's structure learning, and gives an improved learning algorithm.

The rest of the paper is organized as follows. Section 2 reviews related work about BN's structure learning. In Section 3, we introduce BN and its learning methods. In Section 4, we describe the improved learning algorithm of BN in detail. Section 5 reports our experimental results. Finally, we conclude the paper in Section 6.

2 Related work

The development of data mining has stimulated the interest in learning the structure of Bayesian network from data. In general, there are two basic approaches for building its structure. The first one poses the BN's learning as an optimization problem [1, 2, 3, 4]. The second approach poses BN's learning as a constraint satisfaction problem [5, 6].

0-7803-9035-0/05/$20.00 ©2005 IEEE

As each has its own weaknesses, a lot of hybrid algorithms [7, 8, 10] uniting these two approaches have been developed in the last decade. The general idea is quite straightforward. First, the conditional independence(CI) tests are performed to get an initial network that people are willing to consider, which reduce the search space. Then, a search algorithm is called to find a good network structure which has the best motivated score.

Friedman proposed a faster iterative algorithm [10], which restricts the parents of each variable to belong to a small subset of candidates by means of dependence learned from data, then searches for a network that satisfies these constraints. In [7], the researchers introduced a novel hybrid algorithm, which also combines the CI test with the score-and-search process. Based on a similar idea, Qiang Lei [8] proposed a hybrid algorithm, the I-B&B-MDL algorithm, which restricts the search space by using heuristic knowledge and enhances the search efficiency by Branch&Bound technology. Given an order of nodes, the I-B&B-MDL algorithm is an efficient one in comparison with other algorithms. Unfortunately, when the number of nodes is large, there are two major problems for I-B&B-MDL. Since the number of tuples of each conditional set is too large, it is expensive to collect various statistics about the data and to compute the mutual information among the variables even if only performing lower order independence tests. Moreover, without optimizing for search process by heuristic knowledge, there are still many worthless recursive calls.

Aiming at these problem of I-B&B-MDL, in this paper, we give some improved strategies, and propose an improved algorithm called enhanced I-B&B-MDL (EI-B&B-MDL).

3 Learning the structure of Bayesian Network

3.1 Bayesian Network

A BN is represented by $BN = \langle \mathbf{X}, A, \Theta, \rangle$, where $\langle \mathbf{X}, A \rangle$ depicts a directed acyclic graph, each node $X_j \in \mathbf{X}$ represents a domain variable. Each arc $a \in A$ represents a probabilistic dependency between the associated nodes. $\Theta = \{\theta_j\}$ is a set of network parameters, where θ_j depicts a conditional probability distribution table associated with each node X_j, which quantifies how much a node depends on its parents. In graph way, a BN specifies a unique joint probability distribution over $\mathbf{X}$:

$$P(X_1, X_2, \cdots, X_n) = \prod_{j=1}^{n} P(X_j | \Pi(X_j)) \qquad (1)$$

Where n is the number of variables in set $\mathbf{X}$, $\Pi(X_j) = \{X_k : k \in \phi(j)\}$ denotes the set of parents of X_j in the graph, $\phi(j) = \{1, 2, \cdots, j-1\}$ is the sequence number set of parent nodes.

3.2 Conditional Independence Test

In BN's learning, a CI test is a typical metric that checks the independence relationship between two variables under conditional set of variables. The basic of CI is the metric of information flow in information theory, thus the mutual information of two variables X_1, X_2 is defined as

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}, \qquad (2)$$

and conditional mutual information is defined as

$$I(X_i, X_j | C) = \sum_{x_i, x_j, c} P(x_i, x_j) \log \frac{P(x_i, x_j | c)}{P(x_i | c)P(x_j | c)} \qquad (3)$$

where C is a conditional set of nodes, P denotes the instance frequency observed from sample database. The mutual information can show if the two variables are dependent and if so, how close their relationship is. Hence, when $I(X_i, X_j | C)$ is smaller than a certain threshold value ε, we can say that X_i is independent of X_j given the set C, or else X_i is dependent of X_j. So we can deduce if there is a connection between two variables in light of the mutual information [6]. A χ^2 test is another method to estimate a connection between two variables [5]. Given a degree of confidence σ, we can deduce if there is a connection between two variables using the p-value generated by χ^2 test. In effect, if the p-value is greater than or equal to σ, X_i is independent of X_j, this implies that there is no direct connection between these two nodes. Otherwise, if the p-value is lesser than σ, X_i is dependent of X_j, this implies that a connection between X_i and X_j can exist in the resultant network.

3.3 The MDL-based Learning Algorithm

In a feasible solution space, the MDL-based learning algorithm searches for an optimal structure that satisfies the condition of minimum description length (MDL). The basic flow of the algorithm can be described as follows.

If the finite set of random variables for a BN is denoted by $\mathbf{X} = \{X_1, \cdots, X_n\}$, where each variable X_j, $j \in J = (1, 2, \cdots, n)$, may take on values from a finite set $V^j = \{0, 1, \cdots, v^j - 1\}(v^j \geq 2$: some integer). The learning of the network structure virtually is the identifying of parent sets $\{\Pi^1, \Pi^2, \cdots, \Pi^n\}$, where Π^j is the set of nodes that a node X^j depends on. Given a sample set $x^{(i)}$=$\{x^1, \cdots; x^n\}$ of $\mathbf{X}$, $i \in$\{1,2,$\cdots$,N\}, where

N is the sample capacity, then $x^{(i)} \in V = \prod_{j \in J} V^j$, $x^N = x^{(1)}x^{(2)} \cdots x^{(N)} \in V^N$. While G is the set of possible network structures, $g \in G$, then the description length $L(g, x^N)$ of BN is expressed as [9]:

$$L(g, x^N) = H(g, x^N) + \frac{k(g)}{2} \log N \tag{4}$$

where the empirical entropy $H(g, x^N)$ describes the fitness of each possible structure to the observed data, and $H(g, x^N) = \sum_{j \in J} H(j, g, x^N)$, $k(g)$ is the description for the complexity of nodes. It stands for the number of independent conditional probabilities embedded in the structure g, and $k(g) = \sum_{j \in J} k(j, g)$.

$$H(j, g, x^N) = \sum_{s \in S(j,g)} \sum_{q \in V^j} -n[q, s, j, g] \log \frac{n[q, s, j, g]}{n[s, j, g]} \tag{5}$$

$$k(j, g) = (v^j - 1) \prod_{k \in \phi(j)} v^k \tag{6}$$

$$n[s, j, g] = \sum_{i=1}^{N} I(\pi_i^j = s) \tag{7}$$

$$n[q, s, j, g] = \sum_{i=1}^{N} I(x_i^j = q, \pi_i^j = s) \tag{8}$$

where $S(j, g)$ is the instance set of the corresponding parent nodes.

Based on the above preparation [9], the problem of learning BN becomes a search problem for a structure with MDL metric. Generally, an exhaustive recursive search was applied to the MDL-based search procedure. This search examines all possible local changes in the set of parent nodes, so the cost of those evaluations is acute for massive datasets.

3.4 The B&B-MDL-based Learning Algorithm

In order to reduce the computational complexity for empirical entropy, Suzuki proposed a Branch&Bound-MDL-based learning algorithm (B&B-MDL) [9], which can avoid worthless recursive calls for some branches of a search tree by estimating the MDL score. In other words, if the value of MDL_1 in the last step is smaller than some branch's lower bound of MDL_2 in current step, and if the lower bound can be computed with a lower cost, then the further recursive calls in this branch can be avoided, namely, the branch including remanent search nodes can be pruned. Figure 1 shows a instance of B&B-MDL search tree for the set of parents with four nodes. In this tree, there are two pruned branches, and six scoring evaluations can be avoided.

Although the structure complexity of the node increases along with the number of its parent nodes increasing, the

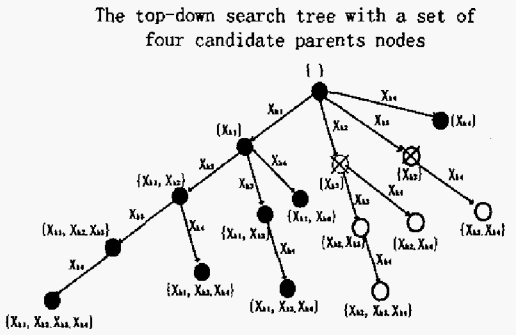

Figure 1. A instance of B&B-MDL search tree for the set of parents with four nodes

value of empirical entroy is nonnegative and descending monotonously, and the decrement of empirical entropy is at most the current empirical entropy $H(j, g, x^N)$. Thus, for a new increasing parent node q, if

$$H(j, g, x^N) \leq \frac{k(j, g)(v^q - 1)}{2} \log N \tag{9}$$

then $MDL_2 \geq MDL_1$ always holds in this step, namely, because the value of $k(j, g)$ is more increased, any recursive calls is meaningless.

3.5 The I-B&B-MDL-based Learning Algorithm

The B&B-MDL-based learning algorithm improved the MDL-based learning algorithm only from the viewpoint of search. However, most of the candidates are considered to be eliminated in advance based on statistical understanding of the domain. Aiming at this problem, Qiang presented an improved algorithm called as I-B&B-MDL [8]. By using a set of lower order independence tests (χ^2 test), the algorithm restricts the search space and enhances the search efficiency. More precisely, the algorithm uses the mutual information to construct initial network, which restricts the possible parents of each node. Thus, instead of having $j - 1$ potential parents for a node, the algorithm only considers k ($k \ll j - 1$) possible parents in each search. Since the search space is significantly restricted, the search performs faster than B&B-MDL.

Unfortunately, when the number of nodes is large, there are two major problems for I-B&B-MDL. Since the number of tuples of each conditional set is too large, it is expensive that the cost of collecting various statistics about data and computing mutual information even if only performing lower order independence tests. Moreover, because there is extra cost of CI tests, the algorithm cannot ensure that there are enough pruned subtrees to make I-B&B-MDL be more efficient than B&B-MDL.

4 EI-B&B-MDL

In order to overcome above drawbacks, we propose an Enhanced I-B&B-MDL algorithm (EI-B&B-MDL). There are two major contributions in this algorithm. Firstly, order-0 and partial order-1 independence tests(only for order-1 unilateral double-connection) are used to obtain an original graph of the network, which reduces the number of independence tests and database passes while effectively limiting the search space. Secondly, by means of the heuristic knowledge of mutual information, sort order for candidate parent nodes increases the cut-offs of B&B search tree and accelerates search process. In order to account distinctly for our algorithm, we give the definition of order-1 unilateral double-connection.

Definition 1. Given an arc between two nodes X_i and X_j in a Bayesian network, if there is another path which is the same direction as the arc, and the path only include an extra node X_k, we call this acyclic subgraph as order-1 unilateral double-connection. The sketch map of 1-order unilateral double-connection is shown as Figure 2.

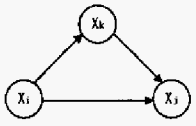

Figure 2. The sketch map of 1-order unilateral double-connection

In that event, we get a set of condition tests $Z'_{ij} = \{X_k | X_i$ and X_j satisfy to order-1 unilateral double-connection, $X_i \prec X_k \prec X_j$, $k \in N$, and $k \neq i, j\}$. Apparently, $Z'_{ij} \subseteq Z_{ij} = \{X_p | p \in N$, and $p \neq i, j\}$, $|Z'_{ij}| \ll |Z_{ij}|$ (Z_{ij} is the condition set of order-1 according to [8]).

The major steps of the learning algorithm are as follows.

Step 1. Given a node ordering, conduct order-0 CI tests for each pair variables in light of Eq. (2), then build the initial graph G_0, in which each arc meets to the constraint condition $I(X_i, X_j) \geq \varepsilon$ (ε is the test threshold), and record the mutual information of each arc in G_0.

Step 2. For every order-1 unilateral double-connection in G_0, conduct order-1 CI tests in light of Eq. (3), and remove the invalid arc that does not pass a CI test. As a result, simplify G_0 to G_1.

Step 3. For each node X_j, ascertain candidate parents $\Pi(X_j)$ of the node according to the structure of G_1, and produce an ordering of parent nodes by sorting each arc's mutual information in ascending order. Then adopt the enhanced B&B-MDL technique to search from top to down, find a $\Pi(X_j)$ with the minimum MDL score and confirm the local optimized structure of X_j. Let $\pi_1 = \phi$, $p_1 = \frac{v^j - 1}{2} \log N$, the main procedure of the EB&B-MDL algorithm is shown as Algorithm 1.

Algorithm 1: EB&B-MDL(π_1, p_1, MDL_1, Π_1)
/* π_1 : the initial set of parents
p_1 : the initial complexity description
MDL_1 : the optimization score
Π_1 : the set of parents after this search */
Begin:
1. Compute the empirical entropy H_1 and $MDL_1 \leftarrow H_1 + p_1$; $\Pi_1 \leftarrow \pi_1$;
2. if $\pi_1 = \Phi$ then $j \leftarrow 0$ else $j \leftarrow$ the last element in π_1;
3. For $j + 1 \leq q \leq k$
/* k: the cardinality of candidate parents' set */
{
$\pi_2 \leftarrow \pi_1 \cup Node(q)$;
/* attach a new node q at the end according to the sort ascending of candidate parents */
$p_2 \leftarrow p_1 \times v^q$;
/* update complexity description of node */
if $H_1 > p_1 \times (v^q - 1)$ then
EB&B-MDL(π_2, p_2, MDL_2, Π_2);
/* predict the MDL of the node, if it diminishs, then call recursive search*/
if $MDL_1 > MDL_2$ then
$MDL_1 \leftarrow MDL_2; \Pi_1 \leftarrow \Pi_2$;
}
End.

---

5 Empirical Study

To evaluate the performance of the proposed algorithm, the benchmark dataset of ALARM network is used. It is a medial diagnostic system containing 37 nodes(variables) and 46 arcs. The Alarm database contains 10000 samples, and each variable has two to four possible values. The platform used for conducting following experiments is a PC with PIV 2.0GHz CPU, 256M memory, and running under Windows XP. The data sets were stored in an access database. When the degree of confidence of χ^2 test is 99.5%, the EI-B&B-MDL algorithm has the same results as [6], which arrives at the best accuracy on same database.

5.1 Experiments on different strategies for independence tests

For different sample capacities, both algorithms are applied, I-B&B-MDL and EI-B&B-MDL, in order to com-

pare their phase results. The results are shown in Table 1. The compression effect of EI-B&B-MDL, which compresses the search space by order-0 and partial order-1 CI tests, is not as good as that of I-B&B-MDL, but as the time saved in CI's testing is by far longer than the time increased in searching (especially when the sample capacity is big), hence the strategy of CI tests can improve the time performance of I-B&B-MDL algorithm.

5.2 Experiments on different methods of sort order

For a search tree with a given topological structure, there are various select order for candidate parent nodes. Different select orders can lead the place of status point in the tree to change, make the predicted evaluation value in every subtree different. As a result, they induce the difference of B&B-MDL algorithms' search efficiency by changing the pruning place and number.

In the worst case (when no subtree is pruned), the B&B search is equal to the exhaustive search. Furthermore, as extra cost of predicted estimate, the time performance of B&B algorithm may increase. Thus it is a key problem for improving search to find a good select order of parents' nodes by means of heuristic knowledge. Table 2 demonstrates that different sort methods result in time performance of the B&B-MDL algorithm under the same conditions of CI tests.

From experimental results, we find that the algorithm's time performance can be primely improved when we sort the parent nodes' order in terms of the ascending order of mutual information and then search with a MDL's scoring. In light of ascending order, there are many subtrees to be pruned , so the algorithm reduces the blindness scoring and searching, enhances the search efficiency. This is the reason that the strategy of sort order is adopted in the EI-B&B-MDL algorithm.

5.3 The Performance of the EI-B&B-MDL Algorithm

Under the same condition, we perform the I-MDL algorithm based on order-0&order-1 CI tests, the I-B&B-MDL algorithm based on order-0&order-1 CI tests, and the EI-B&B-MDL algorithm, respectively. Figure 3 shows the time performance of different algorithms in experiments on currency databases of Alarm. The results show that the running time of our algorithm is lower than that of other algorithms over the whole scope of sample capacity. The advantage is very obvious when the data set is large, namely, the bigger the sample size is, the more obvious the difference is. This is because our algorithm can enhance the search efficiency using heuristic knowledge of mutual information,

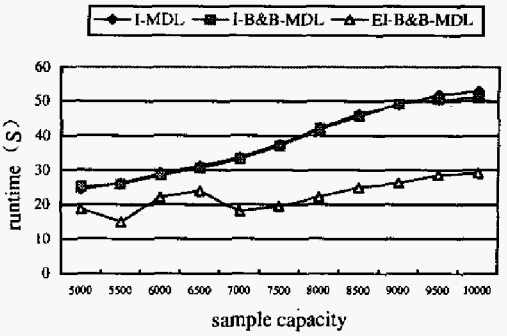

Figure 3. The Comparison of Time Performance of Different Algorithms

and reduce the number of independence tests and database passes by means of a few lower order's CI tests. In a word, the fact that the running time of our algorithm increases so slowly, suggests that our algorithm will be able to handle very large datasets, so the EI-B&B-MDL algorithm is promising.

6 Conclusions

In this paper, we proposed a new improved algorithm, EI-B&B-MDL, for learning Bayesian networks structure efficiently and effectively. The algorithm first makes full use of order-0 and few order-1 CI tests to obtain an original possible graph of the network structure, which reduces the number of independence tests and database passes while effectively restricting the search space. By means of the heuristic knowledge of mutual information, the algorithm fulfills sort order of candidate parent nodes, which increases the cut-offs of B&B search tree and accelerates search process. In experiments on currency datasets of Alarm, the modified algorithm is faster than some hybrid algorithms while keeping with high accuracy, and it is suggested that large data sets can be handled. Hence the modified algorithm is a powerful and efficient algorithm for learning BN.

Our future work includes studying more rigorous bound conditions, finding new heuristic methods for restricting the search space, and developing the application of BN [11].

Acknowledgments

This work is supported by the NSFC major research program: "Basic Theory and Core Techniques of Non-Canonical Knowledge" (60496322, 60496327), Beijing University of Technology Youth Scientific Research Foundation and Doctor Scientific Research Foundation, and

Table 1. The phase results of both algorithms

sample capacity	the number of arcs after order-0 test		the number of arcs after order-1 test		the maximal number of parent nodes		Test time		search time	
	I-B&B-MDL	EI-B&B-MDL	I-B&B-MDL	EI-B&B-MDL	I-B&B-MDL	EI-B&B-MDL	I-B&B-MDL	EI-B&B-MDL	I-B&B-MDL	EI-B&B-MDL
5000	282	282	61	120	7	9	21.313	3.968	3.813	14.704
6000	284	284	61	119	7	9	23.906	4.625	4.485	17.406
7000	286	286	61	121	7	9	28.110	5.407	5.141	12.625
8000	287	287	64	127	7	10	33.516	6.187	8.500	15.953
9000	289	289	65	123	8	13	38.047	6.969	11.187	19.187
10000	293	293	66	127	7	13	41.593	7.797	9.532	21.312

Table 2. The Comparison of Time Performance of Different sort methods

sample capacity	given node order	reverse node order	sort descending by MI	sort ascending by MI
5000	26.046	29.156	31.438	18.672
6000	21.329	40.282	35.860	22.031
7000	25.407	46.891	48.798	18.032
8000	28.453	69.719	65.969	22.140
9000	44.939	105.907	162.751	26.156
10000	51.063	123.546	198.375	29.109

Open Foundation of Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology.

References

[1] D. Heckerman. A tutorial on learning Bayesian networks. Technical report, Microsoft Research, Advanced Technology Division, 1995.

[2] E. Herskovits and G. Cooper. A Bayesian method for the induction of probabilistic networks from data. Machine Learning [J], 1992,9(4): 309-347.

[3] J. Suzuki. Learning Bayesian Belief Networks Based on the Minimum Description Length Principle: Basic Properties [J]. IEICE Transactions on Fundamentals, 1999, E82 (10): 2237-2245.

[4] M. L. Wong, W. Lam, and K.S. Leung. Using Evolutionary Programming and Minimum Description Length Principle for Data Mining of Bayesian Networks [J], IEEE Transactions on Pattern Analysis and Machine Intelligence 1999, 21 (2) : 174-178.

[5] M Luis, de Campos, and J Huete. A new approach for learning belief networks using independence criteria [J]. International Journal of Approximate reasoning, 2000,24(1): 11 37.

[6] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. Learning belief networks from data: An information theory based approach [J], Artificial Intelligence, 2002,137: 43-90.

[7] M.L. Wong, S.Y. Lee, and K.-S. Leung. A Hybrid Approach to Discover Bayesian Networks from Databases Using Evolutionary Programming. Proc. 2002 IEEE Inter. Conf. on Data Mining (2002) 498-505.

[8] L. Qiang, T.-Y. Xiao, and G.-X. Qiao. An Improved Bayesian Networks Learning Algorithm. Journal of Computer Research and Development, 39(10) (2002) 1221-1226.

[9] J. Suzuki. Learning Bayesian Belief Networks Based on the Minimum Description Length Principle: An Efficient Algorithm Using the B&B Technique. IEICE Transactions on Information and Systems, E82-D(2) (1999) 356-367.

[10] N. Friedman, I. Nachman, and D. Peer. Learning Bayesian Network Structures from Massive Datasets: The Sparse Candidate Algorithm. Proc. the Fifteenth Conf. on Uncertainty in Artificial Intelligence (1999) 206-215.

[11] J.Z. Ji, C.N. Liu, J. Yan, and N. Zhong. Bayesian Networks Structure Learning and Its Application to Personalized Recommendation in a B2C Portal, Proc. of 2004 IEEE/WIC/ACM Inter. Conf. on Web Intelligence, IEEE-CS Press (2004) 179-184.