Bayesian Network Structure Learning Based On Rough Set and Mutual Information

Zuhong Feng[1,a], Xiujuan Gao[1,b] and Long Wang[1,c]

[1]Beifang University of Nationalities Yinchuan 750021,China

[a]fengzh_1@163.com, [b] 449086157@qq.com, [c]719535838@qq.com

Keywords: Rough set; mutual information; Bayesian network; structure learning

Abstract. In Bayesian network structure learning for incomplete data set, a common problem is too many attributes causing low efficiency and high computation complexity. In this paper, an algorithm of attribute reduction based on rough set is introduced. The algorithm can effectively reduce the dimension of attributes and quickly determine the network structure using mutual information for Bayesian network structure learning.

Introduction

Bayesian network is a network that expresses the dependent and independent relationships among random variables[1]. We can confirm which class that an unknown sequence belongs to by calculating its largest probability. This method has a certain predictability and is used in many fields as a classifier. Bayesian network learning consists of two parts: structure learning and parameter learning. Structure learning is the core of it. The process of structure learning is very complicate. Especially for those incomplete data sets with high dimension attributes, the   calculation of their structure learning is very large. In this paper, we introduced a method of attribute reduction for rough set to remove the redundant attributes of incomplete data set before structure learning, which can greatly reduce the amount of computation in structure learning. Experiments show the algorithm is effective.

Rough Set Theory

Rough set theory was presented for boundary region by Polish professor Pawlak in the 1980s[2]. This method has strong objectivity due to its mathematical formula. One of the major advantages of rough set theory is that it does not require any prior knowledge. There have been a lot of research results on the theory and application in rough set.   The applications of rough set can be mainly divided into two areas: no decision analysis and decision-making analysis. This paper uses no decision analysis to reduce the number of attributes[3].

Classical Theory of Rough Set**.** A set that can be formed by the union of two elementary sets is called an accurate set, otherwise it is called a rough set[4]. Each rough set can be represented approximately by two related accurate sets, which can be upper approximation or lower approximation. Lower approximation is formed by the union of all elementary sets that are contained by the rough set. Upper approximation is formed by the union of elementary sets that have non-empty intersection with the rough set[2]. Given an information system S=(U,A,V,f)，for any subset $X \subseteq U$,we can use two accurate sets, lower approximation set and upper approximation set, to describe it as follows in Eq. 1:

$$\underline{A}(X) = \cup \{Y_i \mid Y_i \in U \mid ind(A), Y_i \subseteq X\}$$
$$\overline{A}(X) = \cup \{Y_i \mid Y_i \in U \mid ind(A), Y_i \cap X \neq \varnothing\}$$

(1)

One of the core content of rough set theory is the attribute reduction. Some attributes of information system are redundant. The purpose of attribute reduction is to delete those unimportant attributes under the condition of keeping knowledge classification ability unchanged so that the complexity of the database structure can be simplified.

Rough Set Theory And Attribute Reduction Of Incomplete Information Systems.In real life, the widespread presence of incomplete information system makes the equivalence relation no longer exist[5]. So the classical rough set theory is expanded, equivalence relation is relaxed to compatible relation or similarity relation. On the basis of this, we can redefine the lower approximation sets and upper approximation sets.

In the tolerance relation proposed by M. Kryszkiewicz, one of the main concepts is to give a NULL value for the elements which do not have any values in an information table. NULL is a value that can be any value.

Given an information system S=(U,A,V,f), for a subset, B ⊆ A, that has missing attribute values "*"，tolerance relation[3] is defined as follows in Eq. 2:

$$T_B(x, y) \Leftrightarrow \forall a \in B(a(x) = a(y) \vee a(x) = * \vee a(y) = *), \forall x, y \in U$$

(2)

Based on tolerance relation, pseudo code of the attribute reduction algorithm is given below:

```
Input: data set D;
Output：the data set after attribute reduction R;
/*Calling the data discretization program */
Discretize(D)
/*the tolerance matrix based on compatible relation*/
FOR i FROM 1 To n /*n is the number of attributes，m is the number of instances */
    FOR p FROM 1 To m
        FOR q FROM 1 To m
            IF D(p,i)=='*'|D(q,i)=='*'|D(p,i)==D(q,i)
                M(q,p)=1;
            Else
                M(q,p)=0;
            END IF;
        END FOR;
    END FOR
END FOR
M(i);    /*tolerance matrixs M1,M2…Mn*/
/*the process of attribute reduction*/
M(A)=M1&M2…&Mn;
FOR i FROM 1 To n
    IF  M(A)=M(A-i)
        R=D-i; /*i is the unimportant attribute, it can be deleted from D*/
    END IF;
END FOR;
```

Algorithms for Bayesian Network Structure Learning

The algorithm for Bayesian Networks Structure Learning has two categories: one is based on searching and scoring, the other is based on condition independence tests[6]. The process of the first algorithm is relatively simple but usually needs to know the order of nodes since the search space is big. Moreover, it is easy to fall into the local optimal structure when using local or random search according to the decomposability of evaluation function. Therefore, this algorithm has a low learning efficiency. The algorithm based on condition independence tests is more complex and expensive in computation. The algorithm builds network by analyzing the dependencies contained in data set. Under some assumptions, the efficiency of the algorithm can be higher and the global optimal structure can be obtained too.

In this paper we present a learning algorithm using Bayesian network structure learning based on mutual information.The algorithm takes a common database table as input[7]. Every attribute in the database table is regarded as a random variable, and expressed as a node in a Bayesian network. The learning algorithm consists three parts: drafting ,thickening, orienting[8].
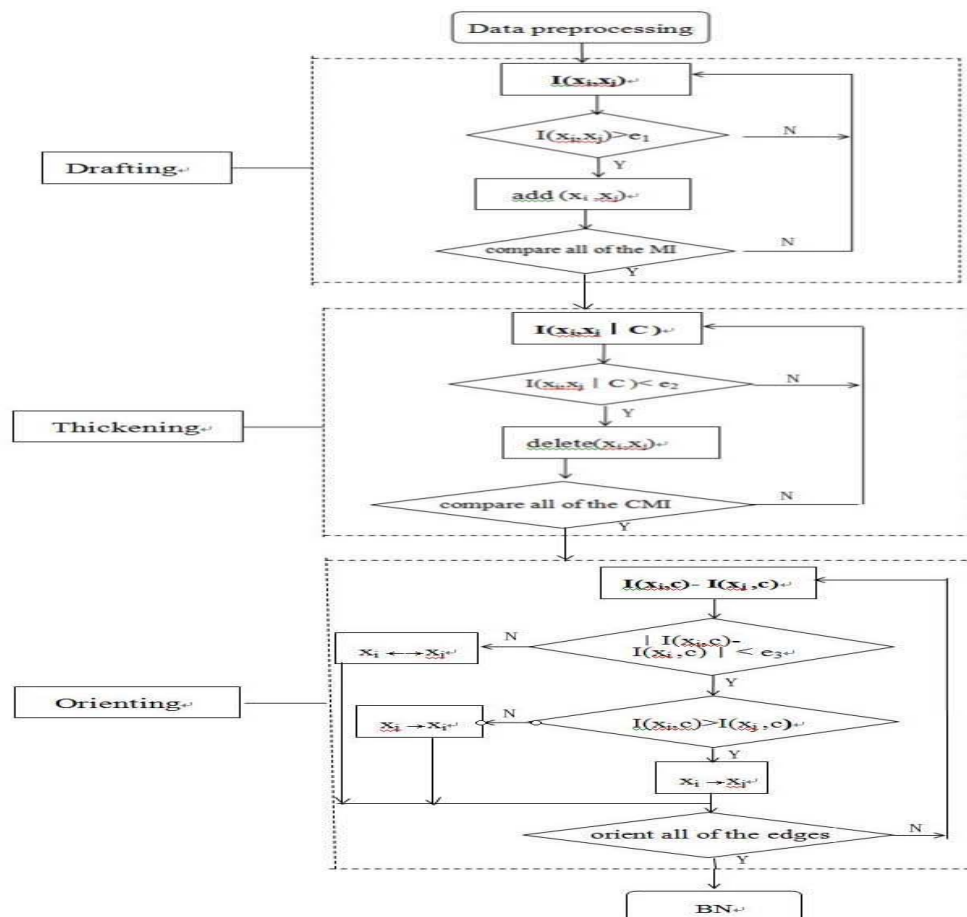The process of algorithm is shown in Fig. 1:

Fig. 1 Bayesian Network Structure Learning

In Fig.1, $I(x_i, x_j)$ and $I(x_i, x_j \mid C)$ represent the mutual information and conditional mutual information between two nodes. They show that the two nodes are interdependent, and how closely their relationship is. The mutual information between the two nodes is defined in Eq. 3:

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \tag{3}$$

The conditional mutual information is defined in Eq. 4 :

$$I(X_i, X_j \mid C) = \sum_{x_i, x_j, c} P(x_i, x_j, c) \log \frac{P(x_i, x_j \mid c)}{P(x_i \mid c)P(x_j \mid c)} \tag{4}$$

where C is a node set. We say that X_i and Y_j is edge independent when $I(X_i,X_j)$ is less than a threshold value ξ and X_i and Y_j is conditional independence under the condition of C when $I(X_i,X_j|C)$ is less than a threshold value ξ.

Experimental process

The Hepatitis Domain data set used in this paper has some missing data. It contains twenty attributes: one of them is a class attribute, the others are non class attributes. The class attribute has no missing data. The non class has some attributes missing. The loss rate is 6%. Various attributes of data set are shown in Table 1.

Table 1. Attributes of data set

sequence	attribute	type	sequence	attribute	type
1	age	continuous	11	spiders	discrete
2	sex	discrete	12	ascites	discrete
3	steroid	discrete	13	varices	discrete
4	antivrals	discrete	14	bilirubin	continuous
5	fatigue	discrete	15	alk phosphate	continuous
6	malaise	discrete	16	sgot	continuous
7	anorexia	discrete	17	albumin	continuous
8	liver big	discrete	18	protime	continuous
9	liver firm	discrete	19	histology	continuous
10	Spleen palpable	discrete	20	class	continuous

First of all, part of the data set is discretized by columns, then the method ofattribute reduction mentioned before is used to extract its feature properties.The algorithm is realized by matlab programming. The experimental result reduced twelve attributes so there are only seven non class attributes left (sex, steroid,liver firm ,varices ,bilirubin,alk phosphate,sgot，numbered from 1 to 7),and one class (numbered 8). Part of the data set after attribute reduction is shown in   Table 2, where "-" means missing data.

Table 2. Part of the data set after reduction

sex	steroid	Liver firm	varices	bilirubin	Alk phosphate	sgot	class
2	2	2	9	2	-	7	2
2	2	2	9	1	6	0	2
1	2	1	16	-	-	-	1
2	2	2	5	2	-	-	2
2	2	2	11	1	-	1	2
2	2	2	8	2	-	4	2

First, fill the feature attributes with missing data respectively. We use statistics method in the experiment. The missing items are filled by the data with the largest probability. Then the mutual information and mutual information of class conditions for each attribute are calculated according to the mutual information theory mentioned above. By comparing the size of the mutual information, the attributes of steroid and liver firm are removed. A sketch (as shown in Fig. 2) can be drawn according to the mutual information between the attributes and the class:
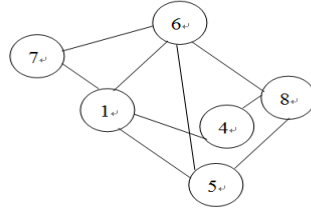

Fig. 2 Drafting

After the edges have been removed according to the conditional mutual information, we can get the network as shown in Fig. 3.
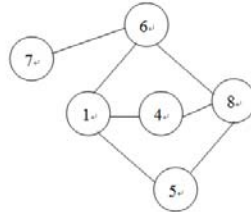

Fig .3 Thickening

Determining the direction of the edge by comparing the mutual information between the class attributes and the non class attributes, we can identify the final network as in Fig.4.
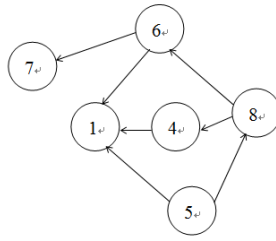
Fig.4 Orienting

Classify testing data set using the Bayesian network without reduction and the Bayesian network with reduction. The test results are shown in Table 3:

Table 3. Compare the results

algorithm	class%		running time (s)
	die	live	
BN	92.31	87.59	23
R-BN	91.86	86.34	11

It can be seen from the experiment that in ensuring same accuracy, the calculation and system overheads are reduced so that the running time is shortened. This shows that the algorithm of attribute reduction for incomplete data set is feasible, and building a Bayesian network using mutual information theory is effective.

Conclusion

In this paper, the attributes of incomplete data set are reduced by using the algorithm of attribute reduction based on rough set. The experiments showed that effectively reducing the dimensions of the attributes and then using mutual information to study Bayesian network structure can reduce the computational complexity and increase the operation efficiency. But the algorithm used in the paper indicates that null value is any value and are potential equal. This may cause two individual objects to be considered to be in the same class even they don't have clearly same or have few same known attribute information. So we should improve the algorithm through increasing the effectiveness and accuracy of the reduction. When using mutual information to study the network structure, missing items are populated with the highest probability values. This would have certain influence on the effect of learning, which should be another direction of improvement in the future.

References

[1] LianWen Zhang,HaiPeng Guo.Introduction to the bayesian network[M].Beijing:Science Press. (2006).

[2] K.S.Chin,Jiye liang,Chuangyin Dang.Rough set analysis algorithms for incomplete information systems,G.Wang et al.(Eds.),Rough Set,Fuzzy Sets,Data Mining,and Granular Computing.2003:264-268.

[3] KaiPing Jang.Attributes Reduction And Rules Mining Under Incomplete Information Systems[D].Jiangxi Normal University,(2009).

[4]Ronan D and Shen Qiang.Learning Bayesian equivalence classes with ant colony optimization[J].Journal of artificial intelligence research,2009,35:391-447.

[5] David Maxwell.Learning equivalence classes of Bayesian - network structures[J].Machine Learning,2002(2):445-498.

[6] Kim SY,Imoto S,Miyano S. Dynamic Bayesian Network and Nonparametric Regression for Nonlinear modeling of Gene Networks from Time Series Gene Expression Data[A]. Computational Methods in Systems Biology,First Intenational Workshop[C].Roverto,Italy, 2003.104-113.

[7] Wang Yue,Tan Shuqiu,Liu Yahui.Network Structural Learning Algorithm Based on Mutual Information[J].COMPUTER ENGINEERING,2011,37(7):62-64.

[8] Jie Cheng,David A.Bell,Weiru Liu,et al.Learning belief networks from data: an information theory based approach[C]. In Proceedings of the Sixth ACM Inter-national Conference on Information and Knowledge Management,2002,325-331.