Bayesian Network Structure Learning and Its Applications in intrusion detection

Feng Zu-hong,Ye Chen and Gao Xiu-juan
Beifang University of Nationalities
Yinchuan 750021, China
fengzh_1@163.com, yechenfish@163.com, 449086157@qq.com

Abstract—**In this paper, an algorithm with attribute reduction based on rough set is introduced. The algorithm can effectively reduce the dimension of attributes and accurately determine the network structure using DBNI with the method of distribution in Bayesian network structure learning with incomplete data. The simulation result shows the algorithm can effectively improve the learning efficiency and detection accuracy of the network.**

Keywords—rough set; mutual information; Bayesian network; structure learning; intrusion detection

I. INTRODUCTION

Bayesian network is a directed acyclic graph [1] [2], which can clearly and visually display the causal relationship between variables. It is a network model of relations among inter-dependent and independent random variables, which can graphically represent the joint probability of random variables. Using probability theory, Bayesian network can handle a variety of uncertain information. . Bayesian network has become an effective tool for knowledge discovery and data mining.

One problem of Bayesian network is that its learning process is complex. The learning process includes two parts: structure learning and parameter learning. The core problem is the structure learning. Especially when data set has high dimensional attributes, the amount of computation by structure learning is very large. Another problem arises too when the data set has missing values (hereinafter referred to incomplete data set). . When the lack of values reaches a certain percentage, it will seriously affect the accuracy of Bayesian network learning and classification.

In this paper, using the attribute reduction algorithm based on rough set [3], redundant attributes of an incomplete data set are reduced before the structure learning process is performed, which reduces the computational complexity of the network structure learning. By proposing DBNI algorithms and taking fully use of valid information included in the incomplete data set, the mutual information and conditional mutual information among attributes are calculated [1] [4] [5]; the attribute nodes that can be interconnected and its direction are determined; a more accurate Bayesian network structure can be obtained. By applying the algorithm to intrusion detection, the simulation experiments show that the algorithm can effectively improve the learning efficiency and detection accuracy of the network.

II. ROUGH SET THEORY AND ATTRIBUTE REDUCTION IN INCOMPLETE INFORMATION SYSTEM

In real life, widespread incomplete information system makes the equivalence relation [7] no longer true. So the classical rough set theory is expanded, equivalence relation is relaxed to compatible relationship and similar relationship. Based on this, upper approximation and lower approximation need to be redefined.

In the tolerance relation proposed by M. Kryszkiewicz [10], the most important idea is to give NULL value to those elements whose values are missing in the information table .NULL is a value that can be any value. By relaxing the equivalence relation, we can use rough set theory in incomplete information system.

Definition 2.1: Given incomplete information system S=(U,A,V,f), for a subset B ⊆ A that has missing attribute values, denote all of its attribute values as "*". A tolerance relation T is defined as below:

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

Denote $T_B(x) = \{y \epsilon U \,|\, xT_By\}$. According to Definition 2.1, the corresponding upper and lower approximate sets can be obtained.

Definition 2.2: Given incomplete information system S=(U,A,V,f), B ⊆ A， X ⊆ U, define the upper and lower approximate sets of X in tolerance relation:

$$\overline{T}_B(X)=\{x \in U \,|\, T_B(x) \cap X \neq \phi\}, \underline{T}_B(X)=\{x \in U \,|\, T_B(x) \subseteq X\}.$$

Definition 2.3: Given incomplete information system S=(U,A,V,f), B ⊆ A , $\forall x \in U$, $T_B(x) = T_A(x)$, and $\forall b \in B$, $T_B(x) \neq T_{B-b}(x)$, then B is called a reduction of A based on tolerance relation.

The pseudocode for the attribute reduction algorithm based on tolerance relation is given below:

Attri_Redu(D,n,m,R):

```
{
Input: data set D;
Output ：data set R after reduction;
/*call data discrete program*/
Discretize(D)
/*find the tolerance matrix based on tolerance relation*/
FOR i FROM 1 To n  /*n is the number of attributes，m is the number of examples of data sets*/
   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);   /* Get n tolerance matrices M1,M2…Mn*/
/* Attribute reduction process*/
M(A)=M1&M2…&Mn;
FOR i FROM 1 To n
IF  M(A)=M(A-i)
       R=D-i; /*i is unnecessary attribute and is deleted from the data set */
     END IF;
END FOR;
}
```

Ⅲ. Bayesian Network Structure Learning Algorithm

Bayesian network structure learning algorithms [2] can be divided into two categories: one is based on scoring and searching, and the other is based on conditional independence testing. In the former method the process is relatively simple but it usually needs to know the order of nodes because the search space is large. it is easy to fall into local optimum structure thus its learning efficiency is low. In the method based on conditional independence testing, its process is complex, and the amount of computation is large. But since it builds a network by analyzing the dependencies contained in the data sample, its learning efficient is higher under some assumptions.

3.1 The Basics of Information Theory

Information theory is the mathematical theory of information transmission and information processing. Its mathematical foundation is probability theory. It is now widely used in machine learning, data mining, statistical mechanics and other important areas.

Lemma 3.1[12] : Let X, Y and Z be discrete random variables. $I(X,Y)$ is called the information about X to Y, $I(X,Y|C)$ is called conditional mutual information between X and Y when C is given . Then

(1) $I(X,Y)\geq 0$; (2) $I(X,Y|C)\geq 0$;

In Lemma 3.1, the equivalence of (1) and (2) is true if and only if X and Y are mutually independent.

The random variables X, Y and C in above definitions and Lemmas can be generalized to sets of variables.

In this paper, we use the Bayesian network structure learning algorithm[9] based on mutual information and above Lemmas. In the algorithm, we take an ordinary database table as input, and each attribute in the database table as a random variable which is expressed as a Bayesian network node. The algorithm is divided into three phases, namely, to build a sketch, to delete edges, and to determine the direction[9]. The BNSL algorithm flowchart is shown in Figure 3-1:

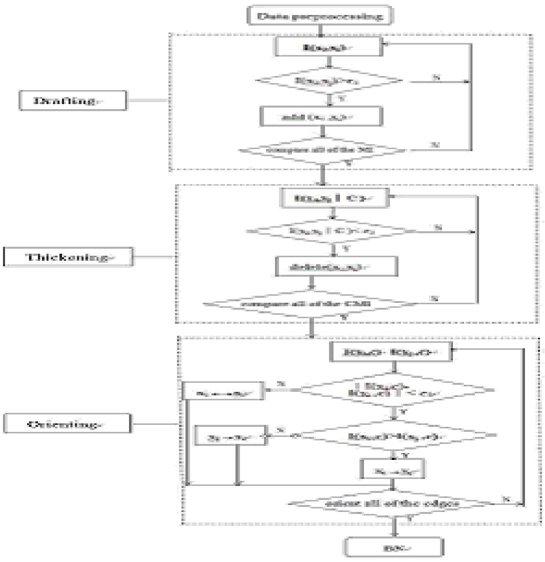

Figure 3-1 BNSL algorithm

In Figure 3-1, I(xi,xj) and I(xi,xj□C) express mutual information and conditional mutual information between two nodes, which indicate if two nodes are mutually independent and the closeness of the relationship between two nodes, and C is a given set of nodes. According to Lemma 2.1, when

$I(X_i, X_j)$ is less than a threshold ξ, Xi and Yj are considered marginal independent. When $I(X_i, X_j \mid C) < \xi$, Xi and Yj are considered conditional independent if C is given.

3.2 Distribution-Based Bayes Network For Incomplete Data -- DBNI

From above, it needs to compute the mutual information and conditional mutual information of random variables for network structure learning [11]. For the problem of the computation of mutual information and conditional mutual information based on incomplete data sets, we propose an algorithm DBNI --Distribution-based Bayes Network for Incomplete data. In order to fully use the information contained in the data set that is not complete, we allocate the number of records that have missing attributes and missing classes to the number of related records that have normal values by the proportion of records that have different normal values according to the number of records that contain the corresponding attributers and classes with different known normal values. The simulation results show that the Bayesian network constructed by this way has higher accuracy.

Let an incomplete data set D have N attributes and class C have L possible values in total. The algorithm DBNI(D, N, C, L) is as follows:

DBNI(D, N, C, L)

Input: incomplete data set D;

Output: mutual information and conditional mutual information, Bayesian Network Structure

1. Count the number of records that nth attributes An takes the jth value a_j^n in D, denoted as f_j^n, and compute frequency $p(a_j^n)$.For any two attributes An、Am, count the number of records that the attribute An takes the jth value a_j^n and the attribute Am, takes the kth value a_k^m, denoted as f_{jk}^{nm}, and compute frequency $p(a_j^n a_k^m)$. Count the number of records that the class attribute C takes the jth value cj, denoted as $f(c_j)$, and compute frequency $p(c_j)$.

Count the number of records that C takes each value cj, denoted as $f(c_j)$,j=1,2, ⋯ ⋯ ,L+1, where $f(c_{(L+1)})$ expresses the number of records that have missing class values.

For each attribute An in D, count the number of records that An takes the jth value a_j^n, denoted as f_j^n, j=1,2,⋯⋯,n_p+1, $f_{(n_p+1)}^n$, expresses the number of records that have missing value An

For any two attributes, count the number of records that An takes the jth value a_j^n and Am takes the kth value a_k^m, denoted as f_{jk}^{nm}, j=1,2,⋯⋯,n_p+1, k=1,2,⋯⋯,m_p+1, where n_p、m_p are the numbers of all possible normal values that An and Am can take respectively. Use $f_{(n_p+1)k}^{nm}$ to express the number of records that a_j^n is missing value but a_k^m takes normal values, $f_{j(m_p+1)}^{nm}$ to express the number of records that a_j^n is normal value but a_k^m is missing, and $f_{(n_p+1)(m_p+1)}^{nm}$ to express the number of records that a_j^n and a_k^m are both missing.

For each attribute An and class attribute C in D, count the number of records that An takes jth value a_j^n and C takes kth value c_k, denoted as f_{jk}^n, j=1,2, ⋯ ⋯ , n_p+1, k=1,2,⋯⋯,L+1, where n_p and L are the number of all possible normal values that An and C can take respectively. Use $f_{(n_p+1)k}^n$, $f_{j(m_p+1)}^n$ $f_{(n_p+1)(L+1)}^n$ to respectively express the number of records for the following three cases: a_j^n is missing but c_k takes normal value, a_j^n is normal value but c_k is missing、and a_j^n and c_k both are missing.

Allocate the number of records that have missing class values to the number of records whose corresponding values are normal by the proportion of records that have different normal values according to the number of records that contain normal values. Let

$c' = \sum_{j=1}^{l} f(c_j)$ Compute $f'(c_j) = f(c_j) + f(c_{l+1}) \times f(c_j) / c'$。

Allocate the number of records that An takes missing values to the number of records whose corresponding values are normal by the proportion of records that have different normal values according to the number of records that contain normal values. Let

$r^n = \sum_{j=1}^{n_p} f_j^n$ Compute : $f_j^{n\prime} = f_j^n + f_{(n_p+1)}^n \times f_j^n / r^n$

Allocate $f_{(n_p+1)k}^{nm}$, $f_{j(m_p+1)}^{nm}$ $f_{(n_p+1)(m_p+1)}^{nm}$ to the number of records that have normal values by the proportion of records that have different normal values according to the number of records that contain normal values. Let

$r_j^n = \sum_{k=1}^{m_p} f_{jk}^{nm}$, $s_k^m = \sum_{j=1}^{n_p} f_{jk}^{nm}$, $s = \sum_{i=1}^{n_p} r_i$,

Compute:

$f_{jk}^{nm\prime} = f_{jk}^{nm} + f_{j(m_p+1)}^{nm} \times r_j^n / s + f_{(n_p+1)k}^{nm} \times s_k^m / s + f_{(n_p+1)(m_p+1)}^{nm} \times f_{jk}^{nm} / s$ Allocate

$f_{(n_p+1)k}^n$, $f_{j(m_p+1)}^n$ $f_{(n_p+1)(L+1)}^n$ to the number of records that have normal values by the proportion of records that have different normal values according to the number of records that contain normal values. Let

$q_j^n = \sum_{k=1}^{L} f_{jk}^n$, $t_k = \sum_{j=1}^{n_p} f_{jk}^n$ $q^n = \sum_{j=1}^{n_p} q_j^n$ Compute : $f_{jk}^{n}{'} = f_{jk}^n + f_{j(L+1)}^n \times q_j^n / q^n + f_{(n_p+1)k}^n \times t_k / q^n + f_{(n_p+1)(L+1)}^n \times f_{jk}^n / q^n$

Use $f'(c_j)$ 、 $f_{jk}^{n}{'}$ and $f_{jk}^{nm}{'}$ to compute probability $p(c_j)$ and joint probability $P(a_j^n c_k)$ and $P(a_j^n a_k^m)$,

$$p(c_j) = \frac{f'(c_j)+1}{\sum_{k=1}^{l} f'(c_k)+L},\quad p(a_j^n) = \frac{f_j^{n}{'}+1}{record+n_p}$$

$$p(a_j^n a_k^m) = \frac{f_{jk}^{nm}{'}+1}{record+n_p m_p},\quad p(a_j^n c_k) = \frac{f_{jk}^{n}{'}+1}{record+n_p L}$$

Where record expresses the total number of records in D.

2. Compute $p(a_j^n a_k^m c_i)$。

Let f_{jki}^{nm} be the number of records that attributes An takes jth value a_j^n, attribute Am takes jth value a_k^m, and class C takes value c_i . In addition, make the following notations:

$f_{(n_p+1)ki}^{nm}$: the number of records that a_j^n is empty but a_k^m and c_i are not ;

$f_{j(m_p+1)i}^{nm}$: the number of records that a_k^m is empty but a_j^n and c_i are not ;

$f_{jk(L+1)}^{nm}$: the number of records that a_j^n and a_k^m is not empty but c_i is ;

$f_{(n_p+1)k(L+1)}^{nm}$: the number of records that a_j^n and c_i are empty but a_k^m is not ;

$f_{j(m_p+1)(L+1)}^{nm}$: the number of records that a_j^n is not empty but a_k^m and c_i are ;

$f_{(n_p+1)(m_p+1)i}^{nm}$: the number of records that a_j^n and a_k^m are empty but c_i is not ;

$f_{(n_p+1)(m_p+1)(L+1)}^{nm}$: the number of records that a_j^n , a_k^m and c_i are all empty.

Allocate $f_{(n_p+1)ki}^{nm}$, $f_{j(m_p+1)i}^{nm}$, $f_{jk(L+1)}^{nm}$, $f_{(n_p+1)k(L+1)}^{nm}$, $f_{j(m_p+1)(L+1)}^{nm}$, $f_{(n_p+1)(m_p+1)i}^{nm}$ and $f_{(n_p+1)(m_p+1)(L+1)}^{nm}$ to the number of records that have normal values by the proportion of records that have different normal values according to the number of records that contain normal values. The computation process is indicated as follows:

First calculate : $r_i^n = \sum_{k=1}^{m_p} \sum_{j=1}^{L} f_{ikj}^{nm}$ $r_k^m = \sum_{i=1}^{n_p} \sum_{j=1}^{L} f_{ikj}^{nm}$, $r_j = \sum_{i=1}^{n_p} \sum_{k=1}^{m_p} f_{ikj}^{nm}$, $r_{ij}^n = \sum_{k=1}^{m_p} f_{ikj}^{nm}$, $r_{kj}^m = \sum_{i=1}^{n_p} f_{ikj}^{nm}$, $r_{ik}^{nm} = \sum_{j=1}^{L} f_{ikj}^{nm}$, $s_d = \sum_{i=1}^{n_p} r_i^n$。 then compute :

$$f_{ikj}^{nm}{'} = f_{ikj}^{nm} + f_{(n_p+1)kj}^{nm} \times r_i^n / s_d + f_{i(m_p+1)j}^{nm} \times r_k^m / s_d + f_{ik(L+1)}^{nm} \times r_j / s_d + f_{(n_p+1)k(L+1)}^{nm} \times r_{ij}^n / s_d + f_{j(m_p+1)(L+1)}^{nm} \times r_{kj}^m / s_d + f_{(n_p+1)(m_p+1)j}^{nm} \times r_{ik}^{nm} / s_d + f_{(n_p+1)(m_p+1)(L+1)}^{nm} \times f_{ikj}^{nm} / s_d$$

Use the number of records computed to solve joint probability $p(a_i^n a^m{}_k c_j)$ $p(a_i^n a^m{}_k c_j) = \dfrac{f_{ikj}^{nm}{'}+1}{record+n_p m_p L}$

Solve $P(a_j^n | c_k)$ and $p(a_i^n a^m{}_k | c_j)$ respectively using $p(c_j)$ and joint probabilities

$P(a_j^n c_k)$ and $p(a_i^n a^m{}_k c_j)$ obtained in 1 and 2 and the following formulas:

$$P(a_j^n | c_k) = \frac{P(a_j^n c_k)}{P(c_k)},\qquad p(a_i^n a^m{}_k | c_j) = \frac{p(a_i^n a^m{}_k c_j)}{p(c_j)}$$

3. Compute mutual information and conditional mutual information among attributes according to the probabilities calculated in 1,2 and 3. Then use the mutual information and conditional mutual information calculated to call BNSL to obtain the Bayesian network structure learning**.**

Ⅳ INTRUSION DETECTION SIMULATION AND ITS RESULTS ANALYSIS

The test data used in our simulation is KDDCUP'1999 datasets, which were coarse-grained pretreated by Columbia University. Since the resulted datasets contain both discrete and continuous data, it needs a further pretreatment before using them. The pretreatment made herein includes: feature selection from the dataset by manual analysis, discretization of continuous data, and numeralization of non-numericaldata.

4.1 Experiment with Complete Datasets And The Result Analysis

First, remove eight attributes that are useless in classification by analysis. These attributes are land, urgent, num_failed_logins, root_shell, su_attempted, num_shells, is_hot_login , num_outbound_cmds. Then perform discretization on continuous data by columns. The details here is to divide the range of variables into several areas so that they can be converted to discrete variables. In our algorithm, we use $k = 1 + 3.32 * \log 2(n)$ to determine the number of areas to divide, where n represents the data amount of the training set. This method was proposed by Mathematician Stallone Gisborne. If numerical attribute value xi is in $[\min + l \times ((\max - \min)/k), \min + (l+1) \times ((\max - \min)/k)$, its discretized value is l, where max is the maximum value and min is the minimum value in the column of attributes, $l = 0,1,...,\lceil k \rceil$.

4.2 Simulation Results and Analysis for Intrusion Detection of Incomplete Data

In this simulation, we still take KDDCUP'1999 data set as our experimental data. We randomly delete 5%、15%、25%、35% and 60% of data pretreated so that cases of a small amount of missing data, a general amount of missing data and a large amount of missing data can be expressed . First, we use the attribute reduction algorithm based on tolerance relation to extract characteristic attributes of data set. According to the DBNI method described in Section 2.2, count the records, carry out the corresponding computation, renew the number of records of corresponding records and compute joint probabilities. After the mutual information and conditional mutual information under a variety of circumstances have been computed, determine the structure of the Bayesian network by using the Bayesian network construction method BNSL. All experiments are implemented using MATLAB programming.

4.2.1 The Experimental Results

When missing data rate is 5%, remove 23 attributes by the attribute reduction algorithm for missing data, then 10 attributes are left. They are count , srv_count , same_srv_rate , diff_srv_rate , srv_diff_host_rate , dst_host_count , dst_host_srv_count , dst_host_same_srv_rate , dst_host_diff_srv_rate , dst_host_same_src_port_rate. Number these attributes as 1 ~ 10. The only class attribute is numbered as 11.

When missing data rate is 15%, remove 19 attributes, then 14 attributes are left. They are protocol_type，service，logged_in , count , srv_count , same_srv_rate , diff_srv_rate , srv_diff_host_rate , dst_host_count , dst_host_srv_count , dst_host_same_srv_rate , dst_host_diff_srv_rate , dst_host_same_src_port_rate , dst_host_srv_diff_host_rate。Number these attributes as 1 ~ 14. The only class attribute is numbered as 15.

When missing data rate is 25%, remove 18 attributes, then 15 attributes are left. They are protocol_type，service，logged_in , count , srv_count , same_srv_rate , diff_srv_rate , srv_diff_host_rate , dst_host_count , dst_host_srv_count , dst_host_same_srv_rate , dst_host_diff_srv_rate , dst_host_same_src_port_rate , dst_host_srv_diff_host_rate , dst_host_srv_rerror_rate. Number these attributes as 1 ~ 15. The only class attribute is numbered as 16.

For the missing data rate15%, the obtained network structures are shown in Figure 4-1, and the corresponding experimental results are shown in Table 4-1.

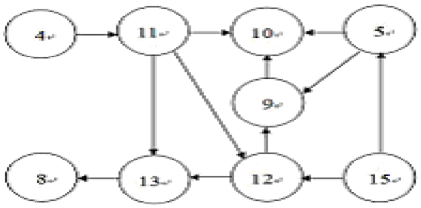

Figure 4-1: Bayesian network structure with missing data rate of 15%

TABLE4-1 THE EXPERIMENTAL RESULTS WITH MISSING DATA RATE OF 15%

Data missing rate	Record type(%)					Run time(s)
	Normal	*Dos*	*Probing*	*R2L*	*U2R*	
15%	*90.22*	*89.75*	*79.83*	*81.24*	*77.81*	*163*

4.2.2 Analysis of the Experiment

By statistical analysis of experimental results, we have the following conclusion.

(1) Analysis of the number of attributes that can be reduced:

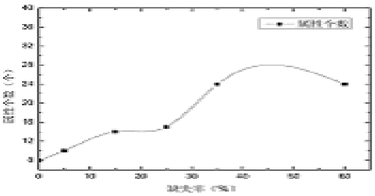

Figure 4-2: Number of attributes under different missing data rates

Figure4-2 shows that missing values are possibly equal to any values since the attribute reduction algorithm used in this paper is based on tolerance relations of rough set. Therefore, the more the data is missing, the less the number of attributes can be reduced, thus the more number of attributes are left.

As shown in Figure 4-2, when data is complete, there are 8 attributes left after reduction. When missing data rate is 5%, 15%, 25%, 35% and 60%, the number of remaining attributes after reduction are 10, 14,15,24, and 24. Therefore, it is believed that when the missing data rate is 5%, the missing data has little effect on attribute reduction; when the missing data rate is 15% -25%, the difference of numbers of attributes that can be deleted is small; when the miss data rate is 35 %-60%, the numbers of attributes that can be deleted is stabilized but the effect is not obvious.

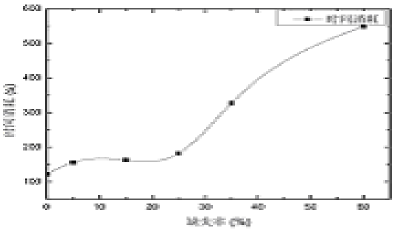

Figure 4-3: The computation time under different missing rates

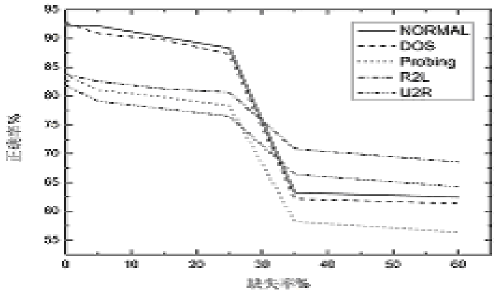

Figure 4-4: Test accuracy rate under different missing rates

(2) Analysis of the relationship between data missing and computation time:

From Figure 4-3, we have the following conclusions: when the missing data rate is within 25%, the difference in the total calculation time is not big, which is in gentle upward trend; when the missing rate is more than 25%, the number of updating the frequencies increases too because of the absence of more data, which leads to more computation time; the computational complexity of the network structure and parameters learning increases too due to the number of attributes that can be reduced, which in turn increases the computational complexity of the detection phase. Therefore, when the data missing rate is greater than 25%, the cost of computing time increases rapidly.

Analysis of the relationship between missing data and detection accuracy

It can be seen from Figure 4-4, with increasing levels of missing data, the detection accuracy gradually declines. But when the missing data rate is within 25%, the detection accuracy only declines about 5% and the rate of decline is relatively flat. When the missing rate is greater than 35%, the detection accuracy rate of the normal case of the 1st class and the 4th class attack is dropped sharply, and the difference of the detection accuracy rates of 35% and 60% is not big, indicating that when missing data rate is greater than 35%, the efficiency of the algorithm starts to fall. When the missing rate at 60%, some detection accuracy rates are even less than 60%, indicating that the algorithm does not apply to cases when large amounts of data is missing.

Through the above experimental results, we can see that the algorithm has better attribute reduction effect for data whose missing rate is less than 25%, which can greatly reduce the computation complexity of the Bayesian network learning and subsequent detection. Moreover, the resulting Bayesian networks through learning have relatively high precision, and have better detection accuracy.

V. Conclusions

In this paper, we have studied the attribute reduction algorithm for incomplete data sets based on rough set theory and Distribution-based Bayes Network for Incomplete data – DBNI, and its application in intrusion detection. The simulation results show that the algorithm has better attribute reduction effect for data whose missing rate is less than 25%, which can greatly reduce the computation complexity. It also shows better detection accuracy in intrusion detection.

Because null values and arbitrary values are potentially equal based on tolerance relation, we perform attribute reduction based on tolerance relation, which affects the results of attribute reduction and detection efficiency to a certain extent. If using other strict relationships to performattribute reduction, it may get better results.

References:

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

[2] Richard E. Neapolitan. Learning Bayesian Networks .Northeastern Illinois UniversityChicago, Illinois.40-43.

[3] Slezak A. Attribute reduction in the Bayesian version of variable precision rough set model[J]. Electronic Notes in Theoretical Computer Science, 2003, 82(4): 1-11.

[4] WANG Wei-ling 1 ,LIU Pei-yu 1 ,CHU Jian-chong. Improved feature selection algorithm with conditional mutual information. Journal of Computer Applications, 2007,27(2):433-435.

[5] Chen X W. Improving Bayesian Network structure learning with mutual information-based node ordering in the k2 algorithm. IEEE Transactions on Knowledge and Data Engineering .May 2008 (vol. 20 no. 5) .628-640.

[6] Hu Q H,ZhaoH,Xie Z X,Yu D R.Consistency based at-tribute reduction. Proceedings of the PAKDD2007 . 2007.

[7] Pawlak Z. Rough set . International Journal of Computer and information Sciences, 1982, 11: 341- 356.

[8] Yao Y Y. On generalizing rough set theory/ / Proceedings of the 9th International Conference on Rough Sets , Fuzzy Set s, Data Mining, and Granular Computing (RSFDGrC 2003) , 2003: 44-51.

[9] Zhang Wen -Xiu, Yao Yi-Yu , Liang Yi. Rough Set and Concept Lattice. Xi'an : Xi'an Jiaotong University Press , 2006 (in Chin ese).

[10] KryszkiewiczM. Rough set approach to incomplete information system s[J]. Information Sciences,1998, 112: 39- 49.

[11] Yager R R.An extension of the naïve Bayesian classifier[J].Information Sciences，2006,176:577-588.

[12] Zhang Lian -Wen, Guo Hai-Peng ,Introduction to Bayesian networks .Sciences Press,2006(in Chinese).