Send Orders for Reprints to reprints@benthamscience.ae

Research on the Data Pre-Processing in the Network Abnormal Intrusion Detection

Xiang Cui[1,2], Guisheng Yin[1,*] and Xuyang Teng[1]

[1]*College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China*

[2]*College of Software Engineering, Harbin University of Science and Technology, Harbin, China*

Abstract: The data pre-processing is a very important step in network abnormal intrusion detection, and directly affects the accuracy of the subsequent detection. In this paper, there are two issues in the network abnormal intrusion detection based on the hierarchical clustering so that some improvements should be made in the data pre-processing stage: first, there is the redundancy and attribute weight problem, each attribute with the weights should be attributing reduced with the use of rough set theory. Second, Aiming to the problem of the continuous data discretization in the rough set theory, an adaptive discrete algorithm for the data characteristics is proposed, and the algorithm determines the intervals of the discretization on the basis of the distribution of the sample attribute values. At last, the two improved methods are experimented and compared with the use of the existing discretization method. The experimental results demonstrate the effectiveness and accuracy of the algorithm.

Keywords: Network abnormal intrusion detection, clustering, data pre-processing, discretization.

1. INTRODUCTION

The network situation awareness has become an important topic of the network field in recent years. The traditional network management cannot meet the requirements so that the future development direction must be based on the network situation awareness whose information is mixed up. The system is a whole concept and the model of the network abnormal intrusion detection is regarded as an important part of the network situation awareness so that the important information can be offered to the sensitive system [1].

The model of the network abnormal intrusion detection can be divided into three steps: the data pre-processing, the building of the abnormal model and the classification of the abnormity. The network abnormal intrusion detection based on the clustering adapts to the typical hierarchical clustering algorithm in the data mining, and different clustering should be divided according to the abnormity with different features in the network so that the abnormal model can be built [2]. Then, the abnormal classification can be done in terms of the abnormal model.

The hierarchical clustering algorithm proposed in the paper adopts the method of the attribute reduction and the weight calculation in the rough set theory to deal with the attribute redundancy H J. The rough set theory is unnecessary to offer other prior information during the procession of the data. The method can calculate the size of each attribute's importance, doing the attribute reduction, removing the redundant attribute, reducing the storage space and improving the operation efficiency [3]. As the basis of the rough set analysis is the discrete data and the data in the network is based on the protocol, time, and various kinds of data in the network flow. The data also includes the numerical attribute and the symbol attribute, the numerical attribute contains the discrete type and the continuous type. Therefore, the continuous data should be effectively and accurately discrete in the data pre-processing stage [4]. Although many people have made wide research to the discrete problem, the problem does not completely belong to the completely generic research topic. In fact, different fields have data features themselves so that it is the fittest to adopt to different discrete methods in terms of the data characteristics. For example, the quantization problem in the image compression requires it has the minimum number of the information after being quantized, while the rough set theory does not have the requirements of the number of the information in the determination table analysis [5].

The two problems in the network abnormal intrusion detection model based on the clustering are improved in the paper.

The obvious disadvantages of the general clustering in the network abnormal intrusion detection model based on the hierarchical clustering are as follows:

1. The clustering does not consider the attribute redundancy;

2. The clustering analysis does not consider the weigh problem of the attributes.

1874-4443/14 2014 Bentham Open

Aiming to the two disadvantages, the attribute reduction in the rough set theory is proposed and the weight can be given in terms of the importance of the attributes [6].

Aiming to the characteristics of the network data in the network abnormal intrusion detection model based on the rough set theory proposes an adaptive discrete algorithm and the problem of a serious imbalance of the different attribute values.

2. THE ATTRIBUTE REDUCTION AND THE WEIGHT CALCULATION METHOD BASED ON THE ROUGH SET THEORY

As to the clustering analysis, a two-dimensional table is formed by the sample set S and the attribute set T Form a two-dimensional table (S,T), and the core of the rough set theory is an information table (as shown in the definition 1). At the point of the model analysis, both have a certain similarity in the structure. The paper regards it as the entry point. The rough set theory introduces the hierarchical clustering based on the distance to do the analysis in the data pre-processing stage of the network abnormal intrusion detection. The following will introduce the clustering analysis on the basis of the rough theory [7].

Definition 1: The representational system of the knowledge: the four group $S=\{U,A,V,f\}$ is a representational system of the knowledge, in which u is the limited set of the objects, A is the limited set of the attributes, that is, $A=C\cup D$. C is the conditional attribute sub sets, D is the determined attribute sub sets. As to any $\alpha\in A$, the $U\to v_a$ of the information reflection always exists $v_a=[l_a,r_a]$ is the attribute. The breaking point can be represented as (a,c) in the value range $v_a=[l_a,r_a]$, in which $\alpha\in A$, c is the real value, any breaking set exists in the $v_a=[l_a,r_a]$, and , f : is the function.

Each attribute of each object is given n information value. If a representational system of the knowledge $K=(\mathrm{U},\mathrm{R})$ is defined to each sub set $\mathrm{X}\subseteq\mathrm{U}$ and an equivalent relationship $\mathrm{R}\subseteq ind(\mathrm{S})$, the two sub sets can be defined as follows:

$$\underline{R}X=\mathrm{U}\{Y\in U/R\,|\,Y\in X\}$$

$\overline{R}X=\mathrm{U}\{Y\in U/R\,|\,Y\cap X\neq\varnothing\}$ is respectively the lower approximation and the upper approximation of x.

$B\pi_R(\mathrm{X})=\overline{R}X-\underline{R}X$ is called as the J R border range of x.

$pos_R(\mathrm{X})=\underline{R}X$ is called as the R positive domain of x.

$neg_R(\mathrm{X})=\mathrm{U}-\overline{R}X$ is called as the chardonnay negative domain of x.

Definition 2: The attribute reduction: If R is an equivalent relationship family, $r\in R$, and $ind(\mathrm{R})=ind(\mathrm{R}-\{\mathrm{r}\})$ regards r is the possible reduced knowledge in the R, that is, called as the unnecessary part in the R. Otherwise, r is necessary in the R. $ind(\mathrm{R})$ dose the classification results in terms of the attributes, while $ind(\mathrm{R}-\{\mathrm{r}\})$ means the classification results after the attributes are reduced. If the $P=r-\{\mathrm{R}\}$ is independent, P is a reduction of the attribute. In the actual application, a classification is very important to the relationship of another classification. Therefore, a concept of the positive domain of a classification is produced comparing to another classification.

Definition 3: The positive domain: If P and Q is the equivalent relationship in the u, the P positive domain in the Q is $\mathrm{pos}_p(Q)$, that is, $\mathrm{pos}_p(Q)=\underline{U}_{xe/Q}\underline{P}XQ$ is the ET obtained from dividing the attribute P into the determined attribute Q in terms of the conditions.

The importance of i conditioned attribute:

$$R_{c-i}(D)=\mathrm{pos}_{c-i}(D)/\mathrm{pos}_c(D) \tag{1}$$

The larger the importance of the attribute, the larger of the weight value has by combining with the weight formula definition:

$$\mathrm{q}_i=r_c(D)-r_{c-i}(D)/\sum_{i=1}^{n}[r_c(D)-r_{c-1}(D)] \tag{2}$$

The distance formula of the clustering can be obtained from the above analysis. If x_i and x_j is the dimensional value of the sample point i and j, the two sample points distance adapt to the improved Euclidean distance:

$$d_{ij}=\sum_{i=1}\sqrt{q_i(x_i-x_j)^2} \tag{3}$$

The following defined formulas have solved the problems of the attribute redundancy and the weight. The hierarchical clustering to the data sample should be done for each abnormity of the network.

3. THE DESCRIPTION OF THE DISCRETE PROBLEM AND THE PRESENT METHODS

In order to meet the requirements of the rough set analysis, the data should be discrete and the data discretization is a part of the data pre-processing part. The continuous conditional attributes should find out the proper branch points under the condition of not changing the ability of the classification. The continuous conditional attribute space should be divided into the discrete space [8].

3.1. The Classification of the Discrete Technology

The discrete algorithm can be divided into two types, and one of them does not consider or seldom considers the

Table 1. Slope value of each interval.

The Line Number:	1	2	3	4	5	6	7	8
The Slope:	16	16	8	4	2	1	1/2	1/4

particularity of the rough set theory. If the discrete method in other subjects is used in the rough set theory, the efficiency of the discretization is not obvious. Another notices the features of the rough set theory, such as the discrete method based on the Boolean Logic and the rough set theory. The following are the typical discrete algorithms:

The equivalent-width and the equivalent-frequency distribution of effects are the most direct and simple discrete algorithms. Both of them adopt the range number K defined by users to divide the value domain $[X_{min}, X_{max}]$ of the attributes into K distributions and each width of the distribution should be equal. The two methods are easily realized, but both during the distribution ignore the sample's distributed information. The flow data in the network abnormal intrusion detection can make the subsequent classification more difficult.

The discrete algorithm is based on the rough set theory. The representative algorithms are the Naive Scalar algorithm and the improved Semi Naïve Scalar algorithm. Both of them do not need the extra parameters, but they just consider an attribute so that the results obtained from the two algorithms can have a conflict [9].

The greedy algorithm: In order to obtain the minimum number of the breaking points, the greedy algorithm can get the breaking point whose importance is the highest so that it is never considered from the whole priority. It just obtains the partial prior solution in a certain meaning.

3.2. The Adaptive Discrete Algorithm

The adaptive discrete algorithm: the sample u is the set of the m objects in the above definition 1 $S = \{U, A, V, f\}$, and the value domain of the corresponding attribute V is (l_a, r_a). The general steps of the algorithm are as follows:

1) The middle value after the attribute region (l_a, r_a) is ordered with the increasing method should be regarded as the first breaking point, and the m_1 object should be corresponded.

2) The corresponding number of m objects in the attribute region (l_a, v_1) and the corresponding number of $(m - m_1)$ objects in the attribute region (v_1, r_a) should be compared. When $m_1 > (m - m_1)$, the middle value v_2 of the attribute in the attribute region (l_a, v_1) with the use of the step 1 is obtained, and it is regarded as the second breaking point and the corresponding objects m_2. The middle distribution of the attribute region $v_3, v_4, ...v_j$ and the corresponding objects $m_3, m_4, ...m_j$; should be found out by repeating the step ① and ②. When the number of the sample in the divided attribute regions (l_a, v_j) and (v_j, r_{j+1}), the $m_l = (m_i - m_l)$, should be stopped.

In order to describe the algorithms in detail, the picture is drawn to explain it. The (l_a, r_a) and the number of the sample m is respectively regarded as unit 1, and the shaft is regarded as the continuous attribute values which are preparing to divide the samples, in which Y coordinate is regarded as the number of the samples. The attribute regions in the shaft which is increasingly ordered with the use of the above steps into two small regions, and then the number of the two regions $7/8 > 1/8$ should be compared. If the region $(0, 7/8)$ which has a more number of the objects, the corresponding attribute region is $(0, 1/2)$. Then the middle value in the region $(0, 1/2)$ is $1/4$. The above steps should be repeated until the slope value is the same (the number of the objects is the same) so that the division of the breaking points can be stopped. At this time, the number of the objects in the two regions is the same so that the value of the attribute discretization is divided by the distribution of the object attributes, and each slope value is as shown in the Table **1**.

The eighth interval is between $1/2$ and 1, the seventh interval is between $1/4$ and $1/2$, the sixth interval is between $1/8$ and $1/4$, and the first interval is between o and $1/128$. The 8 breaking points in the Y coordinate is the discrete value, apart from the first interval and the second interval. The slope values in other each interval are different.

The divided breaking point of the 8 intervals is in the $1/2$ of the total samples in the x coordinate. The $1/2$ in the region whose slope is larger is obtained until the slope in the first interval and the second interval is the same, that is, the number of the sample in the two regions is the same so that the division of the breaking points are stopped. The adaptive discrete algorithm is adopted to solve the serious imbalance of the samples with different attributes for the distribution of the sample points is very imbalanced in the network abnormal intrusion detection. In this way, the number of the breaking points defined by people can be avoided and the subjectivity can also be avoided. The important features of the rough set theory cannot be lost so that it has objectivity.

Table 2. Attribute importance and attribute weights.

Attribute Importance:	**0.95**	**0.95**	**0.83**	**0.90**	**0.92**	**0.80**	**0.90**	**0.83**	**0.84**
Attribute Weights:	0.05	0.05	0.16	0.09	0.07	0.18	0.09	0.15	0.15

Table 3. Recognition results.

Data Set	**Number of Samples**	**Greedy Algorithm**		**Adaptive Algorithm**	
		Recognition Rate (%)	**Error Recognition Rate (%)**	**Recognition Rate (%)**	**Error Recognition Rate (%)**
Iris	150	95	6	96	5
Kdd99	722	82	10	99	3
Glass	214	69	6	72	7

4. EXPERIMENTAL ANALYSIS AND RESULTS

4.1. Experimental Analysis

The experiments prove the effectiveness of the adaptive discrete algorithm and whether the clustering results after the rough set analysis is more accurate, they finish the algorithm programming experiments based on the matlab platform. The classical kdd99 data set is used as the experimental data, and the kdd99 data is based on the network flow statistic feature of time and the network flow statistic feature of the host. The data has 41 network data features, covering the network features of four aspects: the basic feature of the network connection, the content feature of the network connection, the network flow statistic feature of time and the network flow statistic feature of the host. The four network features reflect the network data features. In order to verify the effects in other continuous data discretization, the Iris data and the glass data is used as the compared data in the paper.

The symbol data changes into the character data.

The data should be standardized. The zscore function in the matlab should be data standardized. The zscore formula is as follows:

$$Zscore(\mathrm{x}) = \frac{x - \bar{x}}{s(\mathrm{x})} = \frac{x - \bar{x}}{\sqrt{\frac{(x-\bar{x})^2}{n}}} \tag{4}$$

All data should be ordered with the attribute increasing method.

Many samples focus on the range -2～4 so that the data distribution is very serious imbalanced.

According to the description of the adaptive discrete algorithm, the experimental data should be divided and the data attribute value is distributed in the range -2.1117～15.9375. The first interval is -2.11～7, the second interval is -2.1l～2, the third interval is-2.11～-0.5, the forth interval is -2.11～-1.75 and the fifth interval is -2.11～-2.38.

When the slope is equal in the forth interval and the fifth interval, the division should be stopped. In this way, the 5 breaking points can be divided.

After the kdd99 data is effectively discrete, the data should be roughly analyzed, and respectively calculated by the above defined formula 1 and the formula 2. In this way, the 9 attributes in the 41 attributes are the necessary attributes, and other attributes are the redundant attributes. The 9 attributes are respectively calculated its importance and the weight values from the number 1 to the number 9, as shown in the Table **2**.

After the attribute weights are obtained, the distance can be calculated in terms of the formula 3. The data can be clustered with the hierarchical clustering and the detailed experimental results are as follows：

4.2. Experimental Results

In order to explain the accuracy of the adaptive discrete algorithm, the experiment does the discrete data and then the weights should be calculated. After being calculated, the data should be clustered. The paper compares the adaptive discrete algorithm and other discrete algorithms; the original condition attribute is represented by FA (Former Attribute). After the attribute reduction, the remaining number of the attributes can be shown as RA (Remain Attribute) and the number of the breaking points is BP (Break Point). The recognition results and the remaining condition attributes are respectively as shown in the Table **3** and the Table **4**.

The several kinds of the algorithms are very effective from the following recognition results. The former two algorithms have the same effect, while they have bad effects in the kdd99 data and the test data, which is not as important as the adaptive discrete algorithm. The adaptive discrete

Table 4. Results of the remaining condition attributes.

Data Set	FA	Greedy Algorithm		Adaptive Algorithm	
		RA	BP	RA	BP
Iris	4	3	5	3	6
Kdd99	41	13	6	9	5
glass	9	7	10	6	15

algorithm has a better effect in the network abnormal intrusion detection, and it also has a better effect in the kdd99 data and the network test data to the remaining number of the attributes. In this way, the number of the space dimension can be reduced, the subsequent calculation can also be reduces so that the adaptive discrete algorithm is very effective in the network abnormal intrusion detection [10].

CONCLUSION

The discrete method is very important for the mechanical learning. In recent years, many discrete methods have been proposed and have a certain influence and theoretical significance, but not all of them can have the positive effect of all fields. The paper mainly researches the data pre-processing in the network abnormal intrusion detection based on the clustering, including the improvements of the disadvantages of the hierarchical clustering with the use of the rough set theory, and then the adaptive discrete algorithm is proposed. When the data is discrete, the dimension of the data space can be reduced so that the possessed storage space is smaller. The experimental results prove that the adaptive discrete algorithm is very effective for the discrete stage in the network abnormal intrusion detection and the subjectivity can be avoided. In this way, the accuracy of improving the abnormal detection and the detection efficiency is the important research work of the next step.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation, China (No. 61272186 and No. 61100007), the Natural Science Foundation of Heilongjiang Province, China (No. F201110), the Fundamental Research Funds for the Central Universities, China (No. HEUCF100608) , and the Foundation of Heilongjiang Postdoctoral, China (No. LBH-Z12068).

REFERENCES

[1] M. Kirlidog, and C. Asukb, "A fraud detection approach with data mining in health insurance," *Procedia - Social and Behavioral Sciences*, vol. 62, pp. 989-994, 2012.

[2] H. M. Koupaie, and S. I. J. Hosseinkhani, "Outlier detection in stream data by machine learning and feature selection methods," *International Journal of Advanced Computer Science and Information Technology*, (IJACSIT), vol. 2, pp. 17-24, 2013.

[3] A. M. Said, D. D. Dominic, and B. B. Samir, "Outlier detection scoring measurements based on frequent pattern technique," *Research Journal of Applied Sciences Engineering and Technology*, vol. 6, no. 8, pp. 1340-134, 2013.

[4] Y. Tao, and D. Pi, "Unifying density-based clustering and outlier detection," In: *Proceedings of the 2nd International Workshop on Knowledge Discovery and Data Mining (WKDD)*, 2009, pp. 644-647.

[5] Y. Zhang, N. Meratnia, and P. J. M. Havinga, "Ensuring high sensor data quality through use of online outlier detection techniques," *International Journal of Sensor Networks*, vol. 7, no. 3, pp. 141-151, 2010.

[6] N. K. Tyagi, A. K. Solanki, and S. Tyagi, "An algorithmic approach to data preprocessing in web usage mining," *International Journal of Information Technology and Knowledge Mangement,* vol. 2, no. 2, pp. 279-283, 2010.

[7] N. K. Tyagi, A. K. Solanki, and M. Wadhwa, " Analysis of server log by web usage mining for website improvement," *International Journal of Computer Science Issues*, (IJCSI), vol. 7, no. 4, p. 17, 2010,

[8] K. Taylor and L. Leidinger, "Ontology-driven complex event processing in heterogeneous sensor networks," In: *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web: Research and Applications*, Heraklion, Crete: Greece, 2011, pp. 285-299.

[9] R. Chetan, and D.V. Ashoka, "Data mining based network intrusion detection system:A database centric approach," In: *International Conference on Computer Communication and Informatics*, 2012, pp. 10-12.

[10] I. Silva, L.A. Guedes, P. Portugal, and F. Vasques, "Reliability and availability evaluation of wireless sensor networks for industrial applications," *Sensors*, vol. 12, no. 1, pp. 806-838, 2012.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014