TELKOMNIKA, Vol. 11, No. 7, July 2013, pp. 3736 ~ 3741
e-ISSN: 2087-278X

A Intrusion Detection Method Based on Neighborhood Rough Set

Ming-Xiang He*[1], Dong-Dong Qiu[2]
[1]College of Information Science and Engineering/SDUST Qingdao, Shandong Province/China
[2]College of Information Science and Engineering/SDUST Qingdao, Shandong Province/China
*Corresponding author, e-mail: hmx0708@163.com, qiudongdongshiji@163.com

Abstract

A well-prepared abstract enables the reader to identify the basic content of a document quickly and accurately, to determine its relevance to their interests, and thus to decide whether to read the document in its entirety. The Abstract should be informative and completely self-explanatory, provide a clear statement of the problem, the proposed approach or solution, and point out major findings and conclusions. The Abstract should be 100 to 150 words in length. The abstract should be written in the past tense. Standard nomenclature should be used and abbreviations should be avoided. No literature should be cited. The keyword list provides the opportunity to add keywords, used by the indexing and abstracting services, in addition to those already present in the title. Judicious use of keywords may increase the ease with which interested parties can locate our article.

Keywords*: intrusion detection, data mining, rough set, neighborhood*

1. Introduction

With the popularization of the network and appearance of the new technologies, the importance and effects to society of network are bigger. As losses to the security of nation, enterprise and individual are more severe, the security of network has become one of the most concerns. Intrusion detection is a kind of security mechanism focusing on dynamical monitoring and preventing or system protection. At present intrusion detection has many models and methods. And the introduction of many technologies such as data mining makes the research of intrusion detection to be a hotspot.

Rough set is a mathematical tool to deal with fuzzy and uncertain knowledge and put forward by Pawlak Z who is a Polish mathematician [1]. Its main thought is that we can get the decisions and classification rules of problems by attribute reduction in the premise of keeping classification ability invariant [2]. Rough set has been used in researches of intrusion detection by some researchers, such as a method of intrusion detection which combines support vector machine with rough set by Yi-Rong Zhang and so on [3] and one applying apriori algorithm for rough set by Hong-Jiang Ma [4]. Those methods discretize the data during the data reprocessing process, which would bring information losses inevitably and then results would be distorted to some extent. To this point, the article puts forward a thesis based on neighborhood rough set to intrusion detection. The thesis would not discretize the data during the data reprocessing process, which results in reducing information losses.

2. Related Theory

Intrusion detection analyses the collected information from some key points of the network or the computer system to find out whether there are behaviors and signs to violate the strategy of safety [5]. Intrusion detection system based on the technology of intrusion detection is used in the network and systems using the network. Basing on the monitored results, different intrusion behaviors will be applied by different safe strategy to reduce the harms to the highest extent.

Attribute reduction derives the decisions and classification rules of problems from reduction of knowledge in the premise of keeping classification ability invariant [6]. There are lots of data in the intrusion detection system. We use attribute reduction to divide those

Received January 15, 2013; Revised April 13, 2013; Accepted April 25, 2013

attributes, which would delete redundant information to form rule library and then to help us make a quicker and accurate decision [7].

Suppose A is a subset of topological space (X,τ) and the point $x \in A$.If set U which alse is the open set exists, $U \in \tau$ and U is a subset of A, the point x is an interior point of A and A is a neighborhood of x. If A is a open (closed) set, then it is called open (closed) neighborhood. The neighborhood is applied for attribute reduction, which would reduce the degree of data decentralization and save data integrity to the highest extent.

3. Attribute Reduction Used in Intrusion Detection

Explaining research chronological, including research design, research procedure (in the form of algorithms, Pseudocode or other), how to test and data acquisition [1], [3]. The description of the course of research should be supported references, so the explanation can be accepted scientifically [2], [4].

The sample data during the intrusion detection process include continuous and discrete attributes. Because rough set can only deal with discrete data, we should discretize the data before attribute reduction and then reduce each of discrete attributive characters. We train the intrusion detection system with those characters to get the practical model, which is shown by Figure 1.

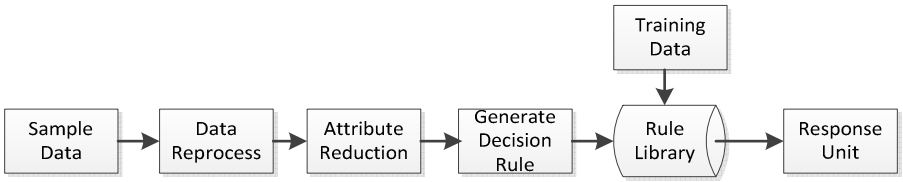

Figure 1. Intrusion detection system based on attribute reduction

With the importance of weighted average of attributes and a null set as starting point, a reduction results from adding the big importance of attribute from all conditional attribute sets gradually. If there are many same importance of attributes, choose any one when adding the big one from surplus conditional attribute sets. The algorithm flow is shown below:

Input: Decision table $DTS = (U, C \cup D, V, f)$, U is record of KDD Cup 1999, C is conditional attribute set and D is decision attribute set [8].

Output: Relative reduction of DTS is called red

Step 1: Assign $red = \phi$

Step 2: Caculate $r = \gamma_c(D)$

Step 3: To $\alpha \in C$, caculate $\sigma_{CD}(\alpha)$

Step 4: Assign $i = 0$, $sig = 0$, $index = INT\_MAX$;

Step 5: Caculate $\overline{\sigma_{redD}}(\alpha_i) = \frac{1}{2}\sigma_{redD}(\alpha_i) + \frac{1}{2}\alpha_{CD}(\alpha_i)$

If $sig < \overline{\sigma_{redD}}(\alpha_i)$, $sig = \overline{\sigma_{redD}}(\alpha_i)$, $index = i$

Step 6: Assign $i = i + 1$, if $i < C.size$, to Step 5;

Step 7: If $sig = 0$, to Step 11;

Step 8: Assign $red = red \cup \{\alpha_{index}\}$

Step 9: $\gamma_{red}(D) = \frac{|POS_{red}(D)|}{|POS_C(D)|}$, if $\gamma_{red}(D) = r$, to Step 11;

Step 10: Assign $C = C - \{\alpha_{index}\}$, to Step 4;

Step 11: End, return red

Attribute reduction resulting from above algorithm is shown below:

(protocol_type, src_bytes, dst_bytes, is_guest_login, dst_host_count, dst_host_same_src_port_rate, dst_host_src_count)

(serviee, src_bytes, dst_bytes, logged_in, dst_host_same_src_port_rate)

(service, duration, src_bytes, dst_host_bytes)

(serviee, duration, src_bytes, dst_bytes)

This algorithm controls the calculation by judging the max importance of weighted average attribute and judges whether it is reduction set by $\gamma_{red}(D)$. Because $POS_C(D)$ is a constant, return the reduction set only when the classified quality is equal with $\gamma_C(D)$.

Suppose the sample data is m and the number of conditional attribute is n in the decision table. At worst, the number of attributes considered every time is n, $n-1$, ..., 1, and total number is $n(n+1)/2$ .The complexity of calculating importance of weighted average every time is approximately $\Theta(m^2)$.Thus, this algorithm would find the satisfied reduction within the complexity $\Theta(n^2m^2)$.

Choose 7136 training and 16561 testing data from the data set of KDD Cup 1999 randomly. In testing data, the number of DoS is 2581, R2L 15, U2R 37 and Probe 79; the number of DoS is 1234 , R2L 12, U2R 20 and Probe 39 for training data.

In order to confirm exprimental results, we assess the capability of intrusion detection system by detection rate and false detection rate. Detection rate (DR) is calculated from the division of detected intrusion data and total intrusion data. False detection rate is defined by the division of normal intrusion data incorrectly and normal one [9].

The process of discretizing data in rough SETES (Rough Set Exploration System) and then calculate training data according to attribute reduction. After reduction, original 41 conditional attributes would leave to 5. Then train the set of the five reduction attributes and classify the training set. And the result of confusion matrix is shown in Table 1.

Table 1. Result set of Rough SETES

	Normal	DoS	R2L	Probe	U2R	Detection Rate
Normal	12874	2	5	23	0	99.78%
DoS	4	2456	0	0	2	99.76%
R2L	0	1	12	0	0	92.31%
Probe	2	1	0	74	4	91.36%
U2R	2	0	0	0	31	93.94%
False Detection Rate	0.06%	0.16%	29.41%	23.71%	16.22%	

Higher detection rate would result from picking up reduction attribute set by attribute detection. But discretizing continuous data before attribute reduction would bring some information losses.

3. Neighborhood Rough Set Used in Intrusion Detection

The algorithm of neighborhood rough set used in intrusion detection applies neighborhood for rough set and would not discretize the data during the data reprocessing process to reduce information losses.

3.1. Data Reprocessing

In order to calculate easily, the first step is data reprocessing. To symbol attributes, discrete attributes are numbered one by one with the set of natural numbers. To numeric ones, continuous are normalized between 0 and 1. The normalization uses equation as below:

$$V_{ij} = \frac{V_{ij} - \min V_j}{\max V_j - \min V_j}$$

V_{ij} means the j attribute value of the i sample. $\max V_j$和$\min V_j$ differently express max and min of all j values.

After normalization, all numeric attributes range between 0 and 1 which can reduce the effects from different attributes. Because the range of different values is equal or greater than 1, the value of neighborhood range from 0 to 1.

3.2. Attribute Reduction

To attribute reduction of neighborhood rough set, we put forward importance of neighborhood attribute because calculating neighborhood would bring a large calculated amount. The algorithm uses forward search to assure that important attributes would be divided into reduction set [10]. The algorithm of reduction is described as below:

Input: One Neighborhood Data System $NDS = \langle U, C \cup D, V, f, \delta \rangle$, U is the domain of discourse, $R = C \cup D$, C is conditional attribute set and D is decision attribute set.

Output: red is one of relative reduction to NDS.

Step 1: $\forall \alpha \in C$, and calculate neighborhood relation NR by neighborhood funcion δ;

Step 2: Assign $red = \phi$;

Step 3: $\forall \alpha_i \in A - red$,and calculating importance of attributes $\sigma_{redD}(\alpha_i) = \gamma_{red \cup \alpha}(D) - \gamma_{red}(D)$;

Step 4: Choose α_j to satisfy $\sigma_{redD}(\alpha_j) = \max(\sigma_{redD}(\alpha_i))$, and if there are more than such α_j, then choose one;

Step 5: If $\sigma_{redD}(\alpha_j) > 0$, assign $red = red \cup \alpha_j$ and to Step 3;

Step 6: End, and return red.

3.3. Picking Up and Collating Rule

The rule library Rul contains lots of redundant information and repeated rules. To that, we should add collating process [11]:

(1) $\alpha \Rightarrow \beta_1$ and $\alpha \Rightarrow \beta_2$ are same rules and we should delete one whose confidence is lower.

(2) $\alpha_1 \Rightarrow \beta$, $\alpha_2 \Rightarrow \beta$ and $\alpha_1 \subset \alpha_2$, then $\alpha_2 \Rightarrow \beta$ is the redundant rule which should be deleted.

In order to comparison of the test, we use the same data. During calculating neighborhood, δ should be inputed artificially and affects the result of reduction. We get a figure wiht the value of δ from 0 to 1 and amount of reduction set, which is shown as Figure 2.

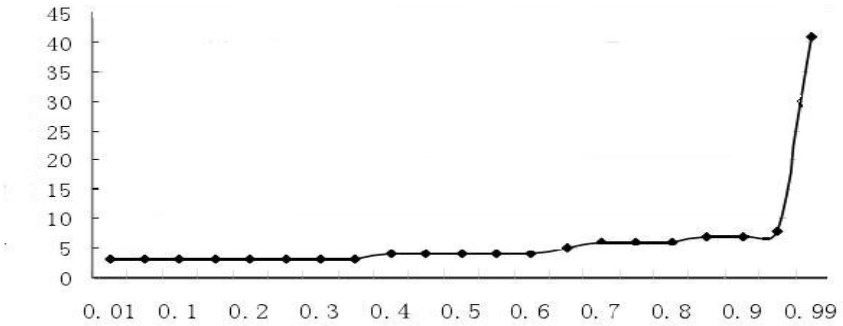

Figure 2. Relation of amount of reduction set and neighborhood value

δ obtains the same amount of reduction set in three ranges: [0.01, 0.3]、[0.45, 0.6]、[0.75, 0.8]. δ is 3、4 and 6 to related range. And δ is closed to 1,reduction set is the all-attribute set. In order to test availiably, we choose δ =0.5 and the result is shown as Table 2:

Table 2. Result of neighbourhood rough set

Title	Detection Rate	False Detection Rate
Normal	99.94%	0.05%
R2L	100%	15.2%
DoS	98.81%	0.06%
U2R	99.98%	5.82%
Probe	100%	6.94%

Comparing rough set after discretizing data, δ is 0.5 in neighborhood rough set. Reduction attribute set is {service, wrong_flagment, dst_host_count, dst_host_same_src_rate} and the set for rough set is {serviee, src_bytes, dst_bytes, logged_in, dst_host_same_src_port_rate}. The result after comparing is shown inTable 3.

Table 3. Comparison of results for rough set and neighbourhood rough set

Algorithm	Amount of Reduction	Detection Rate	False Detection Rate
Rough Set	5	95.43%	0.3%
Neighborhood Rough Set	4	99.75%	0.08%

As shown in the Table 3, the amount of reduction is 4 and detection rate is 99.75% and false detection rate is 0.08% in neighborhood rough set, which is better than rough set.

4. Conclusion

There is still a higher detection rate and lower false detection rate when attribute reduction is used in instruction detection system. But discretizing the data during the data reprocessing process will result in some information losses, which would not reflect the normal information. So this article puts neighborhood to data reprocessing without discretizing the data. With the same testing data and sample data, we can get more important attribute composition by neighborhood rough set. In the comparison of the upper two tests, we can find the advantages of neighborhood attribute reduction used in instruction detection system.

Acknowledgements

This work is supported by the National High Technology Research and Development Program of China (863 Program, No. 2012AA062202) and SDUST CISE Research Fund.

References

[1] Pawlak Z. Vagueness and Uncertainty: A Rough Set Perspective. *Computational Linguistics*. 1995.
[2] Chi-jie Fan, Li-min Chen, Chun-yan Xia. The Approach for Attributes Reduction Based on Rough Set Theory. *Microcomputer Information*. 2010; 26(2-3): 222-228.
[3] Yi-Rong Zhang, Ming Xiao, Shun-Ping Xiao. An Anomaly Intrusion Detection Technique of Support Vector Machine Based on Rough Set Attribute Reduction. *Computer Science*. 2006; 33(6): 64-68.
[4] Hong-Jiang Ma. The Research of Association Rules in Intrusion Detection Based on RST. *Computer Science*. 2006; 33(9): 81-82.
[5] Jing-Yan Li. Instruction detection system and its process. *Network Security Technology & Application*. 2008; 3: 32-34.
[6] Jian-Yuan Wu. Attribute reduction of rough set used in instruction detection. *Journal of Changsha University*. 2010; 24(2): 47-49.

[7] Zong-Hai Xie. One improved method of attribute reduction used on intrucsion detection. *Journal of Mudanjiang Normal University (Natural Sciences Edition)*. 2011; 3: 11-12.
[8] Shang-Zhi Wu. An Algorithm of Attribute Value Reduction and its Application based on Rough Set. *Computer Applications and Software*. 2009; 26(2): 263-265.
[9] Guo Qiang. Ju City: P The Research on Reduction Algorithm of Rough Sets Theory. WuHan. WuHan University of Technology. 2011.
[10] Hu Han, Jian Dang, En-en Ren. Research of Support Vector Classifier Based on Neighborhood Rough Set. *Computer Science*. 2010; 37(2): 229-231.
[11] Wen-li Wang, Li-min Hou. Intrusion detection based on neighborhood rough sets. *Transducer and Microsystem Technologies*. 2010; 29(6): 36-38.