Applied Mechanics and Materials *Online: 2013-08-30*
ISSN: 1662-7482, Vols. 373-375, pp 815-818
doi:10.4028/www.scientific.net/AMM.373-375.815

Rough Set Theory based Hybrid Method for Network Intrusion Detection

Na Jiao[1,a]

[1] Department of Information Science and Technology, East China University of Political Science and Law, Shanghai 201620, P. R. China

[a]zdx.jn@163.com

Keywords: Rough Set Theory, Fuzzy C-Means, Network Security, Intrusion Detection.

Abstract. In this paper, we propose an intrusion detection method that combines rough set theory and Fuzzy C-Means for network intrusion detection. The first step consists of feature selection which is based on rough set theory. The next phase is clustering by using Fuzzy C-Means. Rough set theory is an efficient tool for further reducing redundancy. Fuzzy C-Means allows objects which are belong to several clusters simultaneously, with different degrees of membership. To evaluate the performance of the introduced approach, we applied them to the international Knowledge Discovery and Data mining intrusion detection dataset. In the experimentations, we compare the performance of the rough set theory based hybrid method for network intrusion detection. Experimental results illustrate that our algorithm is accurate model for handling complex attack patterns in large network. And the method can increase the efficiency and reduce the dataset by looking for overlapping categories.

Introduction

Network security has been an important issue since computers have been networked together. Hence the evolution of the Internet becomes increasingly essential for security systems. Intrusions are defined as attempts to compromise confidentiality, integrity, availability of data, or to bypass the security mechanisms of a computer system or network [1].

Intrusion Detection System (IDS) aims at distinguishing an abnormal event from an ordinary one. Intrusion detection techniques can be classified as misuse intrusion detection and anomaly intrusion detection [2]. Misuse intrusion detection refers to detection of intrusions that follow well-known intrusion patterns. It is very popular in detecting the defined attack patterns. Anomaly intrusion detection refers to detection performed by detecting changes in the patterns of utilization or behavior of the system. It is frequently used to detect known and unknown attack. The ideal Intrusion Detection System notifies appropriate person of an attack detection rate along with 0% false positive rate, promptly, with complete diagnosis of the attack.

Generally, there are four categories of attacks [3]:

1) DoS: Denial of Service, for example ping-of-death, teardrop, smurf, SYN flood, and so on.

2) R2L: Unauthorized access from a remote machine, for example guessing password.

3) U2R: Unauthorized access to local super user (root) privileges, for example various “buffer overflow” attacks.

4) Probing: Surveillance and other probing, for example port-scan, ping-sweep, etc..

Some of the attacks (such as DoS, and Probing) may use hundreds of network packets or connections, while on the other hand attacks like U2R and R2L typically use only one or a few connections.

Many approaches have been proposed in intrusion detection. Many data mining approaches, including discovering association rules, have been applied to intrusion detection. Wang and He [4] used the hybrid approach of Rough Set theory and Association Rule Mining to improve the accuracy of intrusion detection.

Rough set theory, proposed by Pawlak [5], is a valid mathematical tool to deal with imprecise, uncertain, and vague information. Researchers have proposed various approaches for feature selection [6-9].

Feature Selection Method Based on Rough Set Theory

Rough Set Theory. A information system is defined as $S=\langle U,A,V,f\rangle$, where U is a non-empty finite set of objects; A is a set of all condition features (also called conditional attributes); $V=\bigcup_{a\in A}V_a$, V_a is a set of feature values of feature a; and $f:U\times A\to V$ is an information function for every $x\in U$, $a\in A$.

For any $B\subseteq A$, an equivalence (indiscernibility) relation induced by B on U is defined as follows:

$$IND(B)=\{(x,y)\in U\times U\mid \forall b\in B, b(x)=b(y)\}. \tag{1}$$

General feature selection algorithm. The general feature selection algorithm starts with the full set, consecutively judges every feature and deletes the features that are superfluous until we obtain a reduction. This method is useful in reducing data set dimensionality.

Algorithm 1. General feature selection algorithm (GFSA)

1. **Set** $Red=A$
2. **do**
3. **for each** $a\in A$
4. **if** $U/IND(P-\{a\})=U/IND(A)$
5. $Red=Red-\{a\}$
6. **loop**
7. **return** Red

Fuzzy C-means Clustering

Fuzzy Set Theory. We briefly describe the Fuzzy C-Means algorithm. The Fuzzy C-Means (FCM) algorithm attempts to partition a finite collection of n elements $X=\{X_1,\cdots,X_n\}$ into a collection of C fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centers $C=\{C_1,\cdots,C_C\}$ and a partition matrix. Fuzzy C-Means algorithm focuses on minimizing the value of an objective function. The objective function measures the quality of the portioning that divides a dataset into C clusters. The objective function is defined as follows:

$$J_m=\sum_{i=1}^{N}\sum_{j=1}^{C}u_{ij}^{m}\left\|x_i-c_j\right\|^2, 1\le m<\infty, \tag{2}$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i–th of d-dimensional measured data, c_j is the d-dimension center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center. A new cluster center value is given below.

$$c_j=\frac{\sum_{i=1}^{N}u_{ij}^{m}\cdot x_i}{\sum_{i=1}^{N}u_{ij}^{m}}, \tag{3}$$

where u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i-th of d-dimensional measured data, and m is any real number greater than 1. The membership of x_i in the cluster j is given below.

$$u_{ij}=\frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i-c_j\right\|}{\left\|x_i-c_k\right\|}\right)^{\frac{2}{m-1}}}, \tag{4}$$

where x_i is the i–th of d-dimensional measured data, c_j is the d-dimension center of the cluster, m is any real number greater than 1, and $\|*\|$ is any norm expressing the similarity between any measured data and the center.

The Fuzzy C-Means algorithm (FCM).

Algorithm 2. The Fuzzy C-Means algorithm (FCM)

Step 1: Given a fixed number C, initialize the cluster center matrix by using a random generator form the original dataset. Record the cluster centers.

Step 2: Initialize the membership matrix $U=[u_{ij}]$, $U(0)$.

Step 3: At k-step: calculate the centers vectors $C(k)=[c_j]$ with $U(k)$.

Step 4: Update $U(k)$, $U(k+1)$.

Step 5: If $\|U(k+1)-U(k)\|<$ threshold then STOP; otherwise return to step 3.

Experiments and Analysis

Experimental Design. The steps in our method as shown in Fig. 1 are:

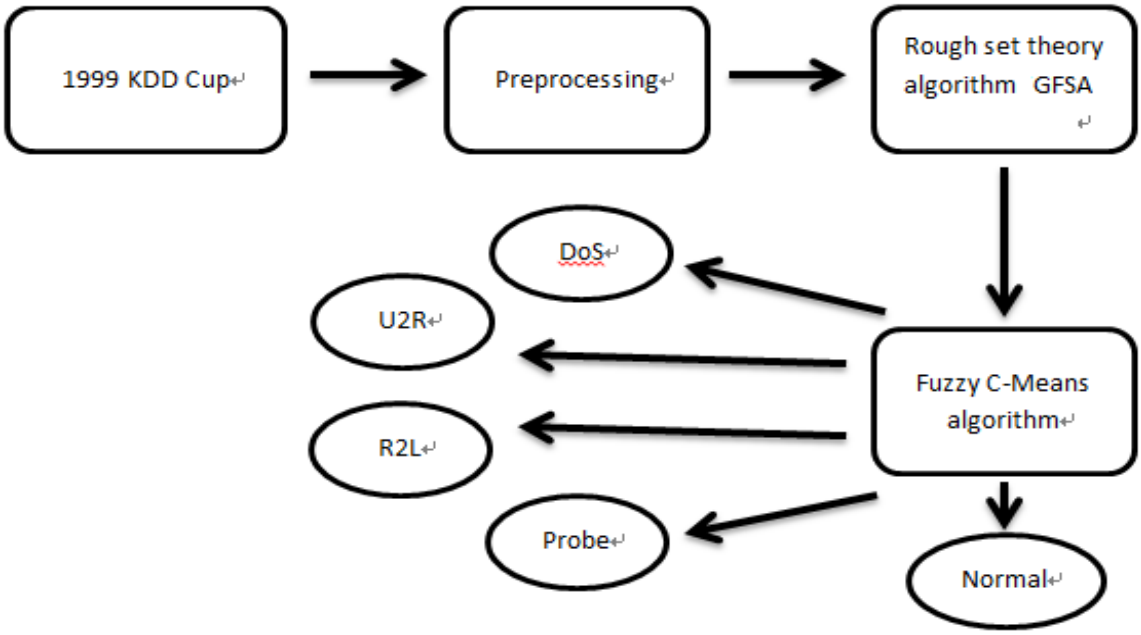

Fig. 1 Overall Intrusion Detection Systems

Experimental Results and Analysis. In the experiments, the dataset consists of 18,216 instances. When general feature selection algorithm (GFSA) was employed, we reduced the dimensionality of the dataset from 42 to 11 features, which were *duration, service, src_bytes, dst_byte, count, srv_count, dst_host_count, dst_host_srv_count, dst_host_diff_srv_rate, dst_host_same_src_port_rate, and dst_host_serror_rate.*

The experimental results are listed in Table 1. As shown in Table 1, the Probe attack type had high detection rate as 100% and low false alarm rate as 0%. The results illustrates that an experimental design which involves rough set theory and Fuzzy C-Means is effective.

Table 2. Experiment results of Fuzzy C-Means with GFSA

Attack Types	Records	Hit	Miss	Detection Rate	False Alarm Rate
Normal	5763	5749	14	99.76%	0.24%
DoS	3530	2897	634	82.05%	17.96%
U2R	70	67	3	95.71%	4.29%
R2L	6689	6145	544	91.87%	8.13%
Probe	2164	2164	0	100%	0%
Summary	18216	17022	1195	93.45%	6.55%

Conclusions

In this paper, we address rough set theory based methods with feature selection to identify subset of features for network security, and use Fuzzy C-Means to intrusion detection. Rough set theory is an efficient mathematical tool for reducing redundancy. Rough set theory based methods can select relevant and informative features. We compare general feature selection algorithm (GFSA) for network intrusion detection. These two methods can increase the efficiency and reduce the dataset. The experimental results demonstrate that the experimental design which involves rough set theory and Fuzzy C-Means is effective.

Acknowledgment

The authors would like to thank the reviewers for their detailed reviews and constructive comments, which have helped improve the quality of this paper. This work was supported by Shanghai University Young Teacher Training Program (no. hdzf10008) and The Research Fund for East China University of Political science and Law (no. 11H2K034).

References

[1] A.O. Adetunmbi and Zhiwei Shi: Proceedings of International Federation for Information Processing (IFIP). Boston: Springer, Vol. 228 (2006), p.525

[2] M. Amini and R.Jalili: Proceedings of the 4th Conference on Engineering of Intelligent Systems (EIS 2004). Madeira, Portugal.

[3] J. Twycross presented at the AISB 2004 Symposium on Immune System and Cognition. Leeds, U.K., (2004)

[4] X. Wang and F. He: IEEE International Conference on Hybrid Information Technology (ICHIT'06). Washington DC, USA: IEEE Press, (2006), p. 114.

[5] Z. Pawlak: International Journal of Computer and Information Sciences, Vol. 11(1982), p. 341.

[6] Leifang Hui, Jiandong Li and Dan Chen: Journal of Beijing University of Posts and Telecommunications, 2011, Vol. 34(2011), p. 30.

[7] Y.Y. Yao and Y. Information Sciences, Vol. 179(2009) p. 867.

[8] Na Jiao, Duoqian Miao and Jie Zhou: Expert System with Applications, Vol. 37(2010), p. 7419.

[9] Na Jiao: Transactions on Rough Sets XIV, Lecture Notes in Computer Science, 6600(2011), p. 100.