http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (18): 4647-4650, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.4647.4650

A New Method of SOM Network Anomaly Detection on the Basis of T-Distribution

Chen Weijun and Jia Weifeng
Anyang Normal University, Anyang, 455000, Henan, China

Abstract: This study introduces a scheme of adaptable distance calculation based on t-distribution which is on the basis of analysis of the scheme of SOM network anomaly detection. The solution sets up a confidence interval between the test sample and BMU distance using t-distribution. It ensures that network anomaly occurs when the distance between the test sample and BMU is not within the range of the confidence interval. In order to test its validity, the improved method is compared with the method of the network anomaly detection based on OC-SVM. Finally, the experimental result shows that this kind of method has characteristics of implementing easily, detecting correctly and having low false alarm rate.

Key words: Network security, anomaly detection, self-organizing map, confidence interval

INTRODUCTION

Intrusion detection has always been a hot and difficult issue in the area of network security. For the moment, most of literatures at home and abroad divide intrusion detection into two types: misuse detection and anomaly detection (Ramadas *et al*., 2003). Misuse detection is a directive intrusion detection method which requires establishing signature database with labeled normal and abnormal date before detection and judges whether the intrusion occurs according to the characteristic agreement of the detected data and labeled data when detecting while anomaly detection only requires establishing a characteristic model of normal data and judges whether the intrusion occurs according to the effective fitting between the present data and the model. The anomaly detection possesses a great value of popularization and engineering application because it needs not to attack and establish database and only to ensure that most of training data are normal.

RELATED RESEARCHES

Self-Organization Map (SOM) Neural Network was first proposed by (Gao *et al*., 2004) which was applied in the voice recognition of Finnish. The SOM learning mechanism which is similar to the human brain is competitive and it is mainly applied in the area of multiple classifications with supervision. In intrusion detection, (Bace, 2000) proposed a scheme of multiclass intrusion detection based on SOM aiming at specific attack types in KDD CUP99 data set and gained a good experimental result which proved the efficiency of SOM neural network in the field of multiclass intrusion detection. In network anomaly detection, Ramadas (kddcup99.html) advanced a scheme of network anomaly detection based on SOM whose basis is that most of the normal data are clustered within a certain distance of BMU, otherwise the anomaly occurs. However, Ramadas's method has two disadvantages: Firstly, the generality is subject to proof because it made experiments just for DNS exploit attack. Secondly, the scope of threshold which is needed to manually input is lack of adaptivity. In view of the above shortcomings, this study checks the validity of the method in KDD CUP data set and comes up with an adaptable solution based on t-distribution which is more convenient for practice in engineering application.

SELF-ORGANIZATION MAP NEURAL NETWORK

Related concepts:

- **Best matching unit (BMU):** For every input vector V in k-dimensions, the Euclid Distance which is between the weights vector of each neuron in SOM and the input vector V, is computed and obtained a set of sequence distance S in which the neuron corresponding to the minimum value is called BMU of V in SOM
- **Nearby neuron:** It usually refers to the neuron nearby BMU in SOM for input vector which can be gained by different methods, such as Bubble Neighborhood Function and Gaussian Neighborhood Function as mentioned in the literature (kddcup99.html)

Corresponding Author: Anyang Normal University, Anyang, 455000, Henan, China

- **Learning function:** The adjusting function of BMU and its nearby neuron weights vector in relation to the input vector with time can be expressed by the equation below:

$$m_i(t+1) = m_i(t) + h_{ci}(t)\left[x(t) - m_i(t)\right]$$

t+1 and t, respectively represent two adjacent time-steps. m_i stands for the weight vector of the ith neuron and x for input vector. $h_{ci}(t)$ is the method of obtaining nearby neuron at time t

SOM training: Do the following for each input vector V_i in the training set:

- Select BMU in SOM of V_i
- Update BMU and the weight vector of its every nearby neuron to make it excitable and suppressive, according to learning function

Repeat the above operation till reaching the intended training times or the change of weight which is smaller than a certain threshold in every learning. The SOM neural network after training can distinguish different types of data. Specific to the field of network anomaly detection, the normal data are usually gathered in a certain distance around BMU in the SOM network trained by using them. Therefore, according to that, we can detect whether or not network anomaly occurs by SOM-BMU distance measure.

IMPROVEMENT OF THE SCHEME OF SOM NETWORK ANOMALY DETECTION

The key of SOM network anomaly detection is how to judge whether or not a distance is beyond the bound of "nearby". Pamadas' method was to estimate artificially an interval range normally distributed beyond which it was considered that anomaly detection occurred. However, it was inconvenient to adjust the parameters owning to measuring interval range artificially first. Aiming at the above shortcoming, this study proposes a scheme of adaptable distance calculation based on t-distribution. The scheme is as follows:

As shown by statistics knowledge, if $X_1, X_2, ..., X_n$ are in a sample of normal population $N(\mu, \sigma^2)$ and $\bar{X}$ and S^2 are, respectively sample mean and sample variance, then:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

is gotten. And according to t-distribution, the feature of y axial symmetry is:

$$P\left\{-t_{\alpha/2}(n-1) \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{\alpha/2}(n-1)\right\} = 1-\alpha$$

Therefore, for the population expectation μ, we can estimate a confidence interval meeting a designed confidence by using sample mean and sample variance. That is to say, the confidence interval that the confidence of μ is $1-\alpha$ will be:

$$\left[\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1), \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\right]$$

Thus setting a quantile, we may get its confidence interval in accordance with the statistical sample. If the measurement (the distance between the test sample and BMU) is within the confidence interval, the network is normal. If not, it is abnormal. The setting of the quantile has character of intuitionisticness which is more humanized. Also, this study introduces slip window operation: the window slips forward a sample if the present sample is detected as normal. It is beneficial to the adaptability of engineering implementation procedure and makes the confidence interval updated regularly.

COMPARISON AND ANALYSIS OF THE EXPERIMENT

In order to validate the efficiency of the method of network anomaly detection based on BMU distance measurement, this study uses KDD CUP99 data set (Kohonen, 2006) for experimenting, meanwhile compares and analyzes the method with One-Class SVM on the same data set. In this section, the content in V.A explains the data set which is used in the writer's experiments and one in V.B shows the detection effect of the method compared and contrasts with that of network anomaly detection based on One-Class SVM.

Experimental data set: Network anomaly detection only needs normal data records in the training set which is because that the character of network anomaly detection firstly requires establishing a behavior model of normal network and judges whether the intrusion occurs according to the fit degree between normal behavior model and the detected sample. This study extracts 2387 items of normal data as training data from kddcup.data_10_percent_corrected in KDD cup99;

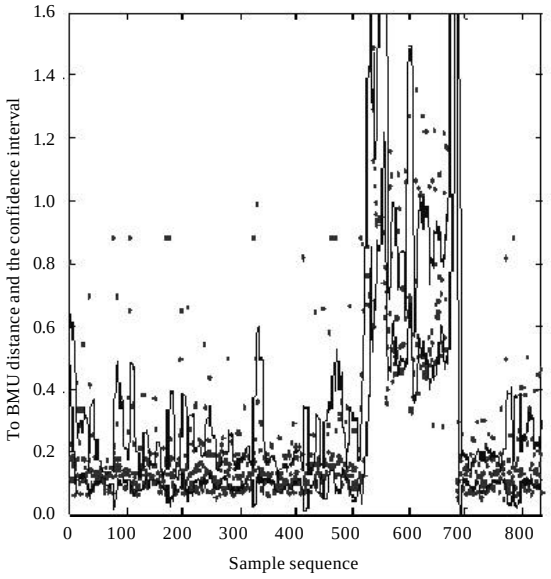

Fig. 1: Normal data and the confidence interval

Table 1: Experimental data set

Data type	Training data set	Test data set
Normal	2387	833
Attack	0	533

extracts 533 items of attack data and 833 items of normal data associations as the test data set of the study for network anomaly detection (Table 1).

Network abnormal detection only detects network abnormal behavior and it doesn't require to do further taxonomic research for abnormal results, so we don't account for the subclass ownership of 533 items of attack data any longer.

Experimental result and contrast: Take confidence level of t-distribution as 0.95, slip window as 10 and deduce the confidence interval of the new detection sample according to the top 10 detection samples. The experimental effect on testing normal data set and attacking data set is shown in Fig. 2 and 3. It is seen from Fig. 2 that the most normal data fall into the confidence interval, so the false rate is in low level. It is seen from Fig. 3 that the most attack data keep off the confidence interval, so this method has high detection rate.

To validate the efficiency of the method, this study gets 5 groups of detection data through adjusting the size of quintile and slip window and then compares 5 groups of detection results with One-Class SVM of different parameters on the same data set (Fig. 3 clearly). It is seen from Fig. 3 that the method of network anomaly detection based on BMU distance measurement with low

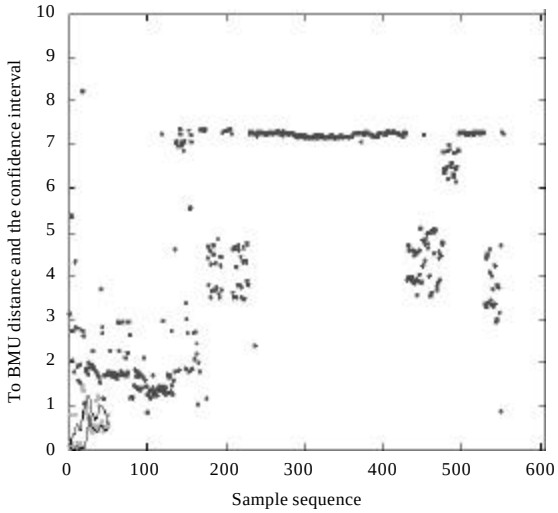

Fig. 2: Attack data and the confidence interval

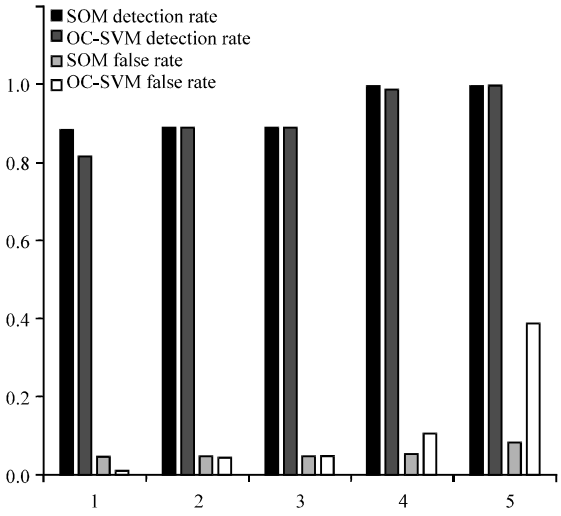

Fig. 3: Experimental results compared with one-class SVM

false rate has higher detection rate than the method of network anomaly detection based on OC-SVM. So, it is effective.

CONCLUSION

Compared with misuse detection, network abnormal detection doesn't need to establish signature data base with attack data it and can detect new attack types. It has become the focus of the field of intrusion detection because of its easy realization from engineering realization angle. The method of network anomaly detection based on BMU distance measurement in this study which can

adjust the confidence interval according to the network situation regularly is more effective with high detection rate and low false rate.

REFERENCES

Bace, R.G., 2000. Intrusion Detection. Macmillan Technical Publishing, Indianapolis, USA.

Gao, J., L. Xu and Y. Dai, 2004. An intrusion detection system model based on Self-organizing map. Proceedings of the 5th World Congress on Intelligent Control and Automation, June 15-19, 2004, Hangzhou, China, pp: 4367-4369.

Kohonen, T., 2006. Self-Organizing Maps. Springer-Verlag, New York.

Ramadas, M., S. Ostermann and B. Tjaden, 2003. Detecting anomalous network traffic with Self-organizing maps. Proceedings of the 6th International Symposium on Recent Advances in Intrusion Detection, September 8-10, 2003, Pittsburgh, PA, USA., pp: 36-54.