Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 3 (2011) 1237–1242

Procedia Computer Science

www.elsevier.com/locate/procedia

WCIT 2010

A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment

Muamer N. Mohammad[a,*], Norrozila Sulaiman[a], Osama Abdulkarim Muhsin[b]

[a]*Faculty of Computer Systems & Software Engineering, University Malaysia Pahang, 26300, Kuantan, Malaysia*
[b]*Collage of Information Technology, University Tenaga Nasional,43000, Selangor ,Malaysia*

Abstract

Nowadays, the using of intelligent data mining approaches to predict intrusion in local area networks has been increasing rapidly. In this paper, an improved approach for Intrusion Detection System (IDS) based on combining data mining and expert system is presented and implemented in WEKA. The taxonomy consists of a classification of the detection principle as well as certain WEKA aspects of the intrusion detection system such as open-source data mining. The combining methods may give better performance of IDS systems, and make the detection more effective. The result of the evaluation of the new design produced a better result in terms of detection efficiency and false alarm rate from the existing problems. This presents useful information in intrusion detection.

Selection and/or peer-review under responsibility of the Guest Editor.

Keywords- Data Mining, intrusion detection system, WEKA.

1. Introduction

Past few years have witnessed a growing recognition of intelligent techniques for the construction of efficient and reliable intrusion detection systems. An intrusion can be defined as "any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource" [1].

Over the years, researchers and designers have used many techniques to design intrusion detection systems. But, there have been one or more problems with present intrusion detection systems. Current anomaly detection methods are mainly classified by statistical anomaly detection [2], detection based on neural network [3] and detection based on data mining [4], etc. Forrest et al firstly use artificial immune method to protect computer [5]. The IDS for the anomaly detection should firstly learn the characteristics of normal activities and abnormal activities, and then the IDS detect traffics that deviate from normal activities. Anomaly detection tries to determine whether deviation from established normal usage patterns can be flagged as intrusions [6]. Anomaly detection techniques is based on the assumption that misuse or intrusive behavior deviates from normal system procedure [7]. The advantage of anomaly detection is that it can detect attacks notwithstanding whether the attacks have been seen before. But the disadvantage of anomaly detection is ineffective in detecting insiders' attacks.

* Corresponding author. Tel.: +6-017-640-9960.
E-mail address: muamer.scis@yahoo.com, norrozila@yahoo.com, osama.alenizi@yahoo.com.

1877-0509
doi:10.1016/j.procs.2010.12.198

In Sodiya [8], a strategy that effectively combined strategies of data mining and expert system was used to design IDS. This technique appeared to be promising, but still with structural and performance problems. Also, combining multiple techniques in designing IDS is a recent event and needs further improvement.

In this paper, the interest work is to improve intrusion system by of combining data mining in WEKA which may give better coverage, and make the detection more effective. The experiments data originated from Computer Lab. The result of the evaluation of the new design produced a better result in terms of the detection efficiency and false alarm rate.

2. Theoretical Background

2.1 Intrusion detection techniques

Intrusion detection techniques can be defined as a system that identifies and deals with the malicious use of computer and network resources. It includes the exterior system intrusion and internal user's non-authorized behavior. It is a technology designed to ensure the computer system security that can discover and inform the non-authorized and abnormal occasions, used to detect the violation of network security. The techniques of intrusion detection can be categorized into two categories [4] anomaly detection and misuse detection.

- *Misuse detection*

Misuse Detection refers to confirming attack incidents by matching features through the attacking feature library. It advances in the high speed of detection and low percentage of false alarm. However, it fails in discovering the non-pre-designated attacks in the feature library, so it cannot detect the numerous new attacks.

- *Anomaly detection*

Anomaly detection refers to storing features of user's usual behaviors into database, then comparing user's current behavior with those in database. If the divergence is huge enough, we can say that there is something abnormal. Its merits lie in it's comparatively irrelevance with the system, its strong versatility and the possibility to detect the attack that has never been detected before. But due to the fact that normal contour conducted cannot gives a complete description of all users' behaviors in the system, moreover each user's behavior changes constantly, its main drawback is the high rate of false alarm.

The system model framework is mainly composed of the following parts: data collecting and preprocessing module, association rule mining module and intrusion detection analysis module, etc as shown in Fig. 1.

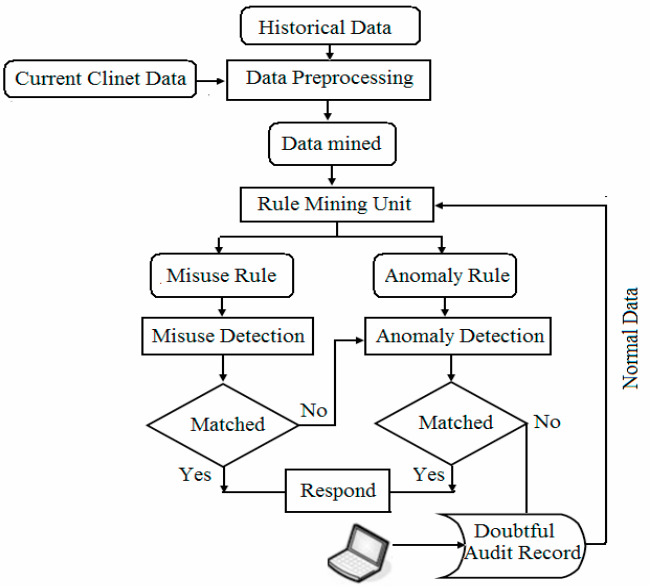

Fig. 1. Data mining system structure in IDS

2.2 Data Mining Technology

Data mining is the latest introduced technology of intrusion detection. Its advantage lies in the fact that it can withdraw the needed and unknown knowledge and regularities from the massive network data and host log data. It is a new attempt to use data mining in achieving network security, both at home and abroad [9,10]. At present, data mining algorithm applied to intrusion detection mainly has four basic patterns: association, sequence, classification and clustering. Data mining technology is advanced for:

- It can process large amount of data.
- It doesn't need the users' subjective evaluation, and is more likely to discover the ignored and hidden information.

Those two are especially applicable to the intrusion detection based on analyzing the abnormality of auditing record [11].

2.3 Weka

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code [12,13]. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA consists of Explorer, Experimenter, Knowledge flow, Simple Command Line Interface, Java interface [14].

The WEKA tool incorporates the following steps [15,16]:

- Analysis and pre-processing of the features in the database and assessing the correctness of the data.
- Definition of the class attributes which divide the set of instances into the appropriate classes.
- Extraction of the potential features to be used for classification.
- Selection of a subset of features to be used in the learning process.
- Investigation of a possible imbalance in the selected data set and how it may be counteracted.
- Selection of a subset of the instances, i.e. the records that learning is to be based on.
- Application of a classifier algorithm for the learning process.
- Decision on a testing method to estimate the performance of the selected algorithm.

3. Proposed System Description

Traditional IDS has some limitations: poor adaptability, inability to detect novel attacks; high modeling cost, slow updating speed; lack of extensibility, etc. Our aim is to design and develop intelligent data mining intrusion detection system and its core part a composite detection engine with anomaly detection and misuse detection features and the two detection engines work serially to detect the user's activity in turn. The system collects the data of database audit system in real time, analyzes the audit data, judges that it is a normal behavior, abnormal behavior or aggressive behavior and responds to the result obtained by the operation behavior and finally reports the result to the manager in a comprehensible form. The model structure is shown as Fig.2.

This part shows the steps that a data mining task is executed on local area network intrusion detection system in Weka software framework. There are four main steps to execute every data mining task with Weka software. In the following steps will illustrate it in detail by a detection task.

1. Initial network data is collected and pretreated as network connection data including particular attributes
2. Network Data packet is generally involved in some important attributes, such as protocol type, destination IP address and flag bit.
3. Subsequently, use association analysis data mining algorithm to handle the connection data and get association rules, thereby obtaining the normal behavior patterns which can be used for abnormal intrusion detection.
4. finally use classification algorithm to carry out rule mining to further distinguish normal behaviors and intrusion behaviors and generate the rules based on misuse detection and meanwhile continue to use analysis data mining algorithm to mine intrusion data sets, extract intrusion patterns, construct an intrusion data feature detection model and update the model according to newly obtained data continuously, used for misuse detection.

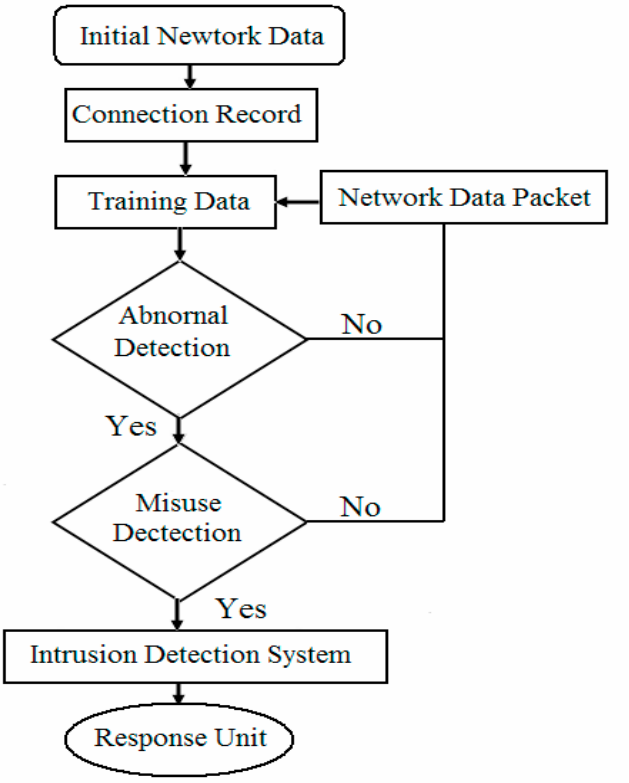

Fig.2 The proposed Data mining system structure in IDS

4. Result and Discussion

Investigate on the performance of proposed algorithm compare with the Signature Apriority algorithm. Nine different-sized databases, from 10Mbytes to 90Mbytes.The experiment were run on a 2.4 GHz Pentium 4 microprocessor with 1 Gbyte RAM. Note that the measurement in this paper of the experimental results is based on the standard metrics for evaluations of intrusion and the ratio between the number of normal connections that are incorrectly misclassified as attacks and the total number of normal connections.

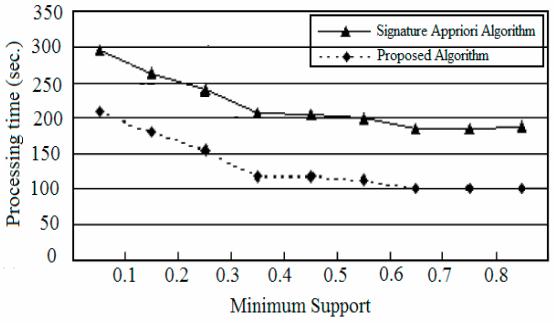

Fig.3 The processing time of Intrusion detection System

Fig. 3 show the minimum support decreases, the processing times of both algorithms increase because of the total number of candidate itemsets increases. It is clear that proposed algorithm is faster than the Signature Apriority no mater what the minimum support is. Therefore, in the real environment, there are not too much candidate itemsets to be generated during each pass of finding signatures.

The influences of minimum support and confidence thresholds on the anomaly detection based on association rule mining are shown as Table 1.

Table1. Influences of Minimum Support and Confidence On Detection Model

No.	Minimum Support (%)	Minimum Confidence (%)	Rule Number	The rule Detect intrusion
1	20	80	279	Yes
2	20	90	151	Yes
3	20	95	124	No
4	25	80	55	No
5	25	90	43	No

Table 1 shows the variation of the minimum support min_sup and the minimum confidence min_conf are relatively small, the cluster rule set can cover all intrusion, at this time, the detection rate will be relatively high, but the rule number is relatively large, which will influence the detection speed.

5. Conclusion

Since the ready-made data mining algorithms is available, intrusion detection based on data mining has developed rapidly. In this paper consider intrusion detection as a process of data analysis by using the predominance of data mining in its effective use of information, this is a method that can automatically generate accurate and applicable intrusion patterns from massive audit data, which makes intrusion detecting system, can be applied to any computer environment. This approach has become a popular topic of research, in the field of inter discipline of network security and artificial intelligence.

References

[1] Heady, R., Luger, G., Maccabe, A., and Servilla, M. The architecture of a network level intrusion detection system. Technical report, Computer Science Department, University of New Mexico, August 1990.

[2] Anderson J. P., et al., “Detecting Unusual Program Behavior Using the Statistical Components of NIDES”, Computer Science Laboratory SRI-CSL-95-06, 1995.

[3] Debar H., Becker M., Siboni D., “A Neural Network Component for an Intrusion Detection System”, Proceedings of IEEE Symposium on Security and Privacy, Oakland CA, pp. 240-251,1992.

[4] Taylor C., Foss J. A., “NATE: Network Analysis of Anomalous Traffic Events, A Low-cost Approach”, Proceedings of New Security Paradigms Workshop, New Mexico USA, pp. 89-96, 2002.

[5] G.W. Dekker, M. Pechenizkiy and J.M. Vleeshouwers, “Predicting Students Drop Out: A Case Study”, Proceedings of 2nd International Conference On Educational Data Mining, Cordoba, Spain, July 1-3, 2009, pp 41-50

[6] Y. Bai and H. Kobayashi, “Intrusion Detection Systems: Technology and Development,” Proceeding of the 17th International Conference on Advanced Information Networking and Applications (AINA’03), 2003.

[7] W. Xuren, H. Famei, and X. Rongsheng, “Modeling Intrusion Detection System by Discovering Association Rule in Rough Set Theory Framework,” International Conference on Computational Intelligence for Modeling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06), 2006.

[8] Sodiya, A.S., Longe, H.O.D. and Akinwale, A.T. (2004), “A new two-tiered strategy to intrusion detection”, Information Management & Computer Security, Vol. 12 No. 1, pp. 27-44.

[9] Ian H. Witten, Eibe Frank "Data Mining: Practical Machine Learning Tools and Techniques (Second Edition) ", Morgan Kaufmann June 2005, 525 pages Paper ISBN 0-12-088407-0.

[10] C. Romero, S. Ventura and E. García, "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, Volume 51, Issue 1, pp. 368-384, 2008, Elsevier Science.

[11] Haimonti Dutta, "Empowering Scientific Discovery by Distributed Data Mining on the Grid Infrastructure". (2007).

[12] D. Patterson, F. Liu, D. Turner, A. Concepcion, and R. Lynch. Performance Comparison of the Data Reduction System. Proceedings of the SPIE Symposium on Defense and Security, Orlando, FL, March 2008.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, "The WEKA Data Mining Software: An Update", ACM SIGKDD Explorations Newsletter, Volume 11 , Issue 1, pp. 10-18, 2009.

[14] Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka

[15] B.X.Wang, D.H.Zhang, J.Wang, et al, "Application of Neural Network to Prediction of Plate Finish Cooling Temperature", Journal of Central South University of Technology, 2008,15(1):136−140.

[16] Ian H.Witten and Elbe Frank, "Datamining Practical Machine Learning Tools and Techniques", Second Edition, Morgan Kaufmann, San Fransisco, 2005.