Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 178 – 182

Procedia Engineering

www.elsevier.com/locate/procedia

Advanced in Control Engineeringand Information Science

Anomaly Detection Based on Enhanced DBScan Algorithm

Zhenguo Chen*,YongFei Li

Department of Computer, North China Institute of Science and Technology, East Yanjiao, Beijing 101601, China

Abstract

As the Internet becomes more pervasive, it is very important that the security mechanisms of a system are designed so as to prevent unauthorized access to system resources and data. The paper proposes a clustering algorithm that exploits enhanced DBScan algorithm in anomaly detection. The algorithm that can be used for mass data processing turns into the hot research point of anomaly detection, to form normal behavior profile on the audit records and adjust the profile timely as the program behavior changed. The experimental result shows that the anomaly detecting based on enhanced DBScan algorithm can a higher detection rate and a low rate of false positives of DARPA data sets.

Selection and/or peer-review under responsibility of [CEIS 2011]

Keywords: DBScan Algorithm; Anomaly Detecting; Intrusion Detection

* Corresponding author. Tel.:+8613473641776
E-mail address: czhenguo@gmail.com

doi:10.1016/j.proeng.2011.08.036

1. Introduction

As the use of computers and network popularly, the network's security of the risks and opportunities have increased. Anderson, while introducing the concept of intrusion detection in 1980 [1], defined an intrusion attempt or a threat to be the potential possibility of a deliberate unauthorized attempt to

- access information,
- manipulate information, or
- render a system unreliable or unusable.

Since then, several techniques for detecting intrusions have been studied.

At present, we can divide the techniques of intrusion detection into two main types. Anomaly detection techniques assume that all intrusive activities are necessarily anomalous. This means that if we could establish a "normal activity profile" for a system, we could, in theory, flag all system states varying from the established profile by statistically significant amounts as intrusion attempts. The concept behind misuse detection schemes is that there are ways to represent attacks in the form of a pattern or a signature so that even variations of the same attack can be detected. This means that these systems are not unlike virus detection systems -- they can detect many or all *known* attack patterns, but they are of little use for as yet unknown attack methods.

Anomaly detection systems are also computationally expensive because of the overhead of keeping track of, and possibly updating several system profile metrics. Anomaly detection commonly used data mining[2] or machine learning methods[3]. Clustering, one method of data mining[4], is paid attention to in the study which is based on unsupervised[5]. ADWICE[6] is a hierarchical clustering [7] to deal effectively with noise of intrusion detection, which can dynamic clustering profile, and stores compressed information using tree structure, thereby reducing the time and space consumption. The algorithm generation system normal action profile by the network data source, with the profile of the monitoring system has achieved a higher detection rate. However, there is a relatively high rate of false positives due to too much noise in the network data.

2. Enhanced DBScan Algorithm

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other cluster.

The DBScan algorithm is a fundamental density-based clustering algorithm. Its advantage is that it can discover clusters with arbitrary shapes. The algorithm typically regards clusters as dense regions of objects in the data space which are separated by regions of low density objects. The algorithm has two input parameters, ε and MinPts. For understanding the process of the algorithm some concepts and definitions has to be introduced.

The DBScan algorithm works as follows. It checks the ε-neighborhood of each object of the dataset, and if in this area more objects than MinPts exist then it is called a core object. Each cluster is grown from a core object by collecting those points that are directly density-reachable from the core point. The algorithm terminates if no more points exist that can be assigned to a cluster. Those objects are treated as noise that could not be assigned to any cluster during the algorithm.

The advantage of this algorithm is that it can find arbitrary shapes of clusters and it exploits the benefit of the natural approach clustering those objects together that forms a dense region. Furthermore it can handle noise as well.

Enhanced DBScan Algorithm, In order to enhance the performance of the DBScan algorithm [8] suggest using R* trees for determining the ε-neighborhood of an object. The R* tree is similar to the B-

tree in a certain manner, namely, in both trees the distance equals between the root and the leaves, and the number of the children of each node is limited. This feature ensures that the processing time on this tree is logarithmic.

For each node belongs an encapsulating region that contains the children of the node. To determine the ε-neighborhood of an object its encapsulating region has to be determined, and the tree has to be traversed from the children of the object to the leaves.

3. Data Set

In this paper, we get information of system calls firstly, and use tf-idf frequency weighting method to pretreatment them. We would make pretreatment process mapping for the text classification process, the relationship between the two is: system calls just like "the word", a process seen as a "document." In dealing with text data, the document as a vector which is made of words, all the data both documents and words component a matrix A.

In the experiment, we extract a large amount of data as training data sample. We only use the frequency of the system calls in the process, and ignore the context of system calls. We will greatly reduce the dimensions of the data in this way and we do not have to consider problem how to reduce the dimension. The difference of characteristic between normal and abnormal behavior become greater through the above data processing methods. And not only makes a higher rate of detection, but also decrease the rate of false positives.

4. Simulation Results

When the normality profile is trained, it is used to detect the unknown data. Solaris operating system of SUN is popular UNIX server operating system, which includes the BSM security module to make Solaris systems to meet TCSEC standards C2-class audit function requirements. BSM security audit subsystem, including the main concepts: the audit log, audit documents, the audit records and audit token and so on. the audit logs is make up of one or more of audit documents, each audit document contains multiple audit records, and each Audit records from a group of audit token (Audit Token) components.

The data we extract if from the DARPA data sets[9], including: a total of 5,046 normal data, from Week 1 of Monday and Tuesday, during the 8:00-16:00. A total of 3,198 abnormal data, from the Week2 Tuesday, Week3 Monday, Week3 Wednesday, Week3 Friday, Week4 Monday. Abnormal data total of 1871 data, from Week6 Thursday.

Table 1: Comparing the number of steps of the different algorithms

Methods	TDR(%)	FAR(%)
k-means	84.1	3.65
Enhanced DBScan	93.7	2.41

To be clearer, we use two performance measures in our experiments. The true detection rate (TDR) measures the percentage of correctly classified in the test set. The false detection rate (FDR) measures the percentage of classified in the test set that are detected false.

According to the experiment results above, the average results which are generated by the two algorithms tested with the same suite of data sets are shown in table 1. It is obviously that we get a higher detection rate and a low rate of false positives of DARPA data sets. So our method is feasible and effective that using in intrusion detection.

5. Summarize

This paper has investigated the use of enhanced DBScan algorithm in the anomaly detection and analyzing the results of the performance measurements, the conclusion can be drawn that the intrusion detection based on enhanced DBScan algorithm achieves the higher recognition accuracy than other method.We get lots of data, and proved the validity of such data through the data pre-processing algorithm, we had done the preparatory work for future after the experiment. And we found an increase clustering algorithm which can dynamic rectify the profile. When we test the algorithm, it has special fast speed.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 60872055), the Natural Science Foundation of Hebei Province (F2010001745) and supported by the Foundation Project of North China Institute of Science And Technology (2008-B-11).

References

[1] J.P. Anderson. Computer security threat monitoring and surveillance. Washington, PA, USA, Technical Report, April 1980.

[2] Varun Chandola, Arindam Banerjee, and Vipin Kumar “Anomaly Detection: A Survey”,August 15, 2007

[3] X. Rao, C.X. Dong and S.Q. Yang, "An Intrusion detection system based on Support Vector Machine". Journal Software,14(4),pp.798-803,2003.

[4] W.Lee, S. Stolfo, K.W.Mok, Adaptive intrusion detection: a data mining approach. Artificial Intelligence Review 14, 2000. pp. 533-567.

[5] L. Portnoy, E. Eskin, S.J. Stolfo, Intrusion detection with unlabeled data using clustering, in: Proceedings of the ACM Workshop on Data Mining Applied to Security, Philadelphia, PA, 2001.pp. 5-8.

[6] K. Burbeck, S. Nadjm-Tehrani. Adaptive real-time anomaly detection with incremental clustering information. Security technical report 12, 2007. , pp. 56 – 67.

[7] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: An Efficient data clustering method for very large databases," Proceedings for the ACM SIGMOD Conference on Management of Data, Montreal, Canada, June 1996. pp. 103-114.

[8] N. Beckmann, HP. Kriegel, R. Schneider, B. Seeger:The R*-tree: An Efficient and Robust Access Method for Points and Rectangles, In Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data, Atlantic Sity, New Jersey, US, 1990. pp. 322-331.

[9] R. Lippmann, J.W. Haines, D.J. Fried, J. Korba, K.Das. The 1999 DARPA off-line intrusion detection evaluation, Computer Networks 34. 2000. pp.579–595.