Global Journal of Computer Science & Technology
Volume 11 Issue 5 Version 1.0 April 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-5861

Intrusion Detection System with Data Mining Approach: A Review

By Madjid Khalilian , Norwati Mustapha , Md Nasir Sulaiman, Ali Mamat
Computer science and information technology University

Abstract- : Despite of growing information technology widely, security has remained one challenging area for computers and networks. Recently many researchers have focused on intrusion detection system based on data mining techniques as an efficient strategy. The main problem in intrusion detection system is accuracy to detect new attacks therefore unsupervised methods should be applied. On the other hand, intrusion in system must be recognized in real-time, although, intrusion detection system is also helpful in off-line status for removing weaknesses of network's security. However, data mining techniques can lead us to discover hidden information from network's log data. In this survey, we try to clarify: first,the different problem definitions with regard to network intrusion detection generally; second, the specific difficulties encountered in this field of research; third, the varying assumptions, heuristics, and intuitions forming the basis of erent approaches; and how several prominent solutions tackle different problems.

Keywords: Data mining, Intrusion Detection, Clustering, Classification.

Classification: *GJCST Classification: FOR Code: 080603,080609,080605*

Strictly as per the compliance and regulations of:

Intrusion Detection System with Data Mining Approach: A Review

Madjid Khalilian , Norwati Mustapha , Md Nasir Sulaiman, Ali Mamat

Abstract- **Despite of growing information technology widely, security has remained one challenging area for computers and networks. Recently many researchers have focused on intrusion detection system based on data mining techniques as an efficient strategy. The main problem in intrusion detection system is accuracy to detect new attacks therefore unsupervised methods should be applied. On the other hand, intrusion in system must be recognized in real-time, although, intrusion detection system is also helpful in off-line status for removing weaknesses of network's security. However, data mining techniques can lead us to discover hidden information from network's log data. In this survey, we try to clarify: first, the different problem definitions with regard to network intrusion detection generally; second, the specific difficulties encountered in this field of research; third, the varying assumptions, heuristics, and intuitions forming the basis of different approaches; and how several prominent solutions tackle different problems.**

IndexTerms- Data mining, Intrusion Detection, Clustering, Classification.

I. Introduction

Nowadays we have many applications with massive amount of data which causes limitation in data storage capacity and processing time. Furthermore, many applications must operate in real-time to achieve theirs objectives. As an important case for these kinds of application, Network Intrusion Detection System (NIDS) can be pointed. Generally we define NIDS as the detection of intrusions or intrusions attempts either manually or via software expert systems that operate on logs or other information available from the system or the network. An intrusion is a deliberate, unauthorized attempt to access or manipulate information or system and to render them unreliable or unusable. If a suspicious activity is from your internal network or system it will also be classified as intrusion. Some popular intrusion as follows:

- Denial of service (DoS): attempts to starve a host of resources needed to function correctly.
- Scan: reconnaissance on the network or a particular host.
- Worms and viruses: replicating on other hosts.
- Compromises: obtain privileged access to a host by known vulnerabilities.

About- Computer science and information technology University Putra Malaysia
E-mails- khalilian@ieee.org,{norwati, nasir, ali}@fsktm.upm.edu.my

Furthermore, we identify some important objectives for IDS as below:

- Detect wide variety of attacks.
- Detect intrusions in timely fashion.
- Present analysis in simple, easy-to-understand format.
- Minimize false positives, false negatives:
  1. False positive: An event, incorrectly identified by the IDS as being an intrusion when none has occurred
  2. False negative: An event that the IDS fails to identify as an intrusion when one has in fact occurred

There are many solutions for intrusion detection that we categorize them into four main groups: anomaly detection, signature based misuse, host based and network based. Many researchers have applied data mining techniques, which are powerful methods for extracting hidden information from huge datasets, for network intrusion detection system. On the other part, traditional data mining is not suitable for this kind of applications so they should be tuned and changed or designed with new algorithms. Besides of speed up and storage capacity, real-life concepts tend to change over time e.g. new attacks should be recognized.

The growth of volume of existing data and insufficiency of data storage capacity lead us to the dynamic processing data and extracting knowledge. The problem is that current IDS are tuned specifically to detect known service level network attacks. At the same time, enough data exists or could be collected to allow network administrators to detect these policy violations. Unfortunately, the data is so enormous, and the analysis process so time-consuming, that the administrators don't have the resources to proactively analyze the data for policy violations, especially in the presence of a high number of false positives that cause them to waste their limited resources.

The nature solution is utilizing data mining techniques. However, data mining can be applied in off-line status. Most previous work focused on off-line environment while on-line system for detecting policy violations is needed.

In next section we addressed general problems in this domain, after that we discuss different solutions in four groups with pros and cons, finally we will have the conclusion.

II. Gaps

- Algorithms suffer from the ability to handle difficult detection tasks without supervision. For example, there is no assumption about types of attacks in data logs but in most methods this parameter should be determined by the expert.
- The algorithms required expert assistant for creating primary model in the form of the number of attacks expected or the expected behavior.
- NIDS is required to re-learn any recurrently occurring patterns for new attacks.
- How to improve the correct rate of intrusion detection.
- How to control the rate of false alarm in anomaly detection.
- Accuracy in terms of detecting concept drift.
- Efficiency in terms of speed is a vital problem in real-time intrusion detection.
- Previous approaches lack precision in detecting outliers.
- Uncertain data: in most applications we don't have sufficient data for statistical operations so new methods are needed to manage uncertain data in accurate and fast fashion.
- In network monitoring, arbitrary value for connection features causes some difficulties in realizing new attacks.
- Data type treatment: different data types (i.e. categorical, ordinal and a mixture of different data types) should be considered in intrusion detection for improving accuracy.
- Alarm Validity: Recent developments in NIDS have heightened the need for determining convenient criteria to validate results. Most outcomes of methods are depended to specific conditions. However, employing suitable criteria in results evaluation is one of the most important challenges in this arena.
- Space limitation: not only time and concept drift are the main complexity in intrusion detection algorithms but also space complexity, where data log would be huge, can be caused difficulties in processing.
- High dimension and scale data: There are high scale data sets with high dimension which should be managed through the processing of data. In huge databases, data complexity can be increased sharply by number of dimensions.
- A Mobile Ad hoc Network consists of a group of autonomous mobile nodes with wireless transmission capability without using any existing infrastructure or centralized administration. The MANET environment is particularly vulnerable due to its dynamic topology, less powerful mobile devices and distributed environment. Most solution was devised for wired network and should modify for MANET.
- Dataset KDD Cup 99 applied in the research is popularly used in current intrusion detection system; however, it is data of 1999, and network technology and attack methods changes greatly, it cannot reflect real network situation nowadays.

III. Traffic Based Methods

If we want to categorize intrusion detection methods, we will recognize two main aspects for grouping approaches, which one group refers to type of attack includes host based and network based. Another group of approaches refers to solutions techniques which are signature based and anomaly detection methods. In continue we review these techniques with their pros and cons.

a) *Host based methods*

This method is based on data source category; consequently, its data comes from the records of various activities of hosts, including system logs, audit operation system information, etc. the main architecture for this kind of methods is similar to network based which is described in the next section. Ref [1] presents a host-based combinatorial method based on k-Means clustering and ID3 decision tree learning algorithms for unsupervised classification of anomalous and normal activities in computer network.

- Advantages
  - Due to the fact that HIDS monitor only the host, it can determined intrude more accurate.
  - It does not need to install extra hardware or software because everything is on the host.
  - Encrypted messages are not serious problem because they received in the host and can be decrypted more easily.

- Disadvantages
  - It can not detect some types of attacks that they need to monitor traffic of network e.g. DOS and DDOS.
  - Redundancy is an important problem in HIDS especially when we want to install this system for a network, because we should have a HIDS for each host.
  - Because of the fact that HIDS should be installed in each host, it is clear that expense of system will be increased.

- Efficiency in terms of speed up is going to be decreased, due to having a monitoring system for each host.

b) *Network based methods*

NIDS is the next main group of methods that is related to source of data. According to the figure 1 some components are realized as follows:

1. Network: it is our environment where it should be controlled against intrusions.
2. Agents: they would be some sensors in the environments by which data must be collected.
3. Network log file(s): it is a repository for storing collected data.
4. Pre-processing: collected data needs to be processed including feature extraction, normalizing, miss value process and so on.
5. Data mining engine: this component is the most prominent component in NIDS based on data mining because creating model take place based on techniques in this part. Classification or clustering would be utilized in this section.
6. Model: after analyzing data from era, a model to identify and prognosticate attack must be created.
7. Human expert: an expert should monitor and control NIDS to take a convenient action.

Advantages and disadvantages stated as follows:

- Advantages
  - Detection of some attacks such as DOS and DDOS need to monitor traffic of whole network and it is possible by NIDS.
  - Low expense is brilliant advantage for NIDS because it is not necessary to install many monitoring systems.

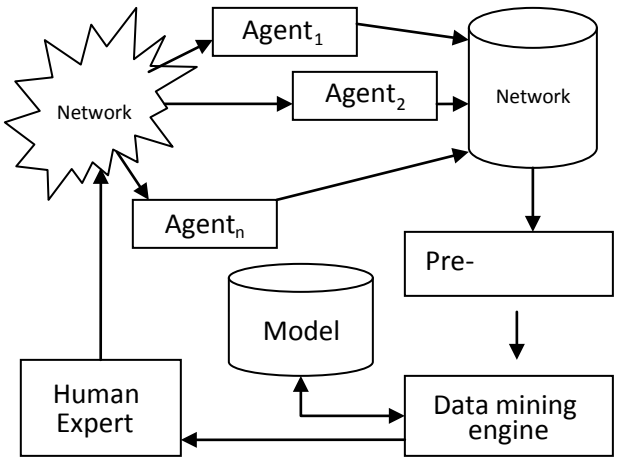

Figure 1 NIDS Based on Data Mining Architecture

- Disadvantages
  - Accuracy is a challenging problem due to losing some data during the process of detection.
  - Encrypted data are problematic in NIDS because of the fact that it is not possible to decrypt data in level of network.
  - In large-scale network more facility is required to monitor network; thus, scalability is another significant problem in NIDS.

IV. Approaches to Solutions

A powerful survey can be found in [2] that it discusses data mining for cyber security applications. For example, anomaly detection techniques could be used to detect unusual patterns and behaviors, Link analysis may be used to trace the viruses to the perpetrators, Classification may be used to group various cyber attacks and then use the profiles to detect an attack when it occurs, Prediction may be used to determine potential future attacks depending in a way on information learnt about terrorists through email and phone conversations. This paper also mentioned about real-time problem in IDS and other challenges include mining unstructured data types.

We divide approaches in two main groups: misuse detection which the main study is the classification algorithms and anomaly detection which the main study is the pattern comparison(association rules and sequence rules) and the cluster algorithms.

a) *Signature-based methods*

The research[3] compares accuracy, detection rate, false alarm rate and accuracy of other attacks under different proportion of normal information. For comparison results of C4.5 and SVM, they demonstrate that C4.5 is superior to SVM in accuracy and detection; in accuracy for Probe, Dos and U2R attacks, C4.5 is also better than SVM; but in false alarm rate, SVM is better. Through test and comparison, the accuracy and detection rate of C4.5 is higher than that of SVM, but false alarm rate of SVM is better. In sampling, the research supposes that the distribution of attack data other than normal data is even, which cannot surely get optimal results, and this should be improved and validated. Another weakness refers to C4.5 parameters that is not optimal, thus the future work should optimize the parameters according to C4.5 parameters and different training dataset. For huge datasets optimizing parameter in SVM takes too much time; however, it is not suitable, for intrusion detection system requires real-timeliness. The future research should aim at the direction where the parameters can be optimized rapidly.

With the concept of field in physics, [4] proposed a data field based method for discrimination

of network behaviors. Similar to electric charge or particle, each data point we concerned has its own influence region and the influence is a function of position giving the force on each point placed at that position. Furthermore, the positive potential and negative potential have been described, by which it can determine the test point's class. This scheme is based on "supervised" learning, whereas unsupervised methods are preferred.

Some advantages and disadvantages are as follows:

- Advantages
  - Specifying exact class of attacks.
  - Efficiency is high and complexity is low.

- disadvantages
  - Many false positives: prone to generating alerts when there is no problem in fact.
  - Cannot detect unknown intrusions.

b) *Anomaly based methods*

The basic idea of clustering analysis originates in the difference between intrusion and normal pattern; consequently, we can put data sets into different categories and detect intrusion by distinguish normal and abnormal behaviors. The common clustering algorithms in data mining include two main categories: hierarchical and partitioning clustering algorithms. Clustering intrusion detection is detection for anomaly with no supervision, and it detects intrusion by training the unmarked data.

Ref [5] considered the outlier factor of clusters for measuring the deviation degree of a cluster. A method has been proposed to compute the cluster radius threshold. The data classification has been performed by an improved nearest neighbor (INN) method. For the unsupervised intrusion detection, they applied a clustering based method that its time complexity is linear with the size of dataset and the number of attributes.

Ref [6] outline a data mining framework for constructing intrusion detection models. To facilitate adaptability and extensibility, they use of meta-learning as a means to construct a combined model that incorporate evidence from multiple base models. They also extend the basic association rules and frequent episodes algorithms to accommodate the special requirements in analyzing audit data. The main shortcoming is lack of devising a mechanical procedure to translate automatically learned detection rules into modules for real-time IDS.

[7] proposed a new weighted support vector clustering algorithm; it can cluster large data set and high-dimensional data effectively. They also introduced the new weighted SVC method to network intrusion detection. The experiments with KDD Cup1999 data demonstrate that proposed method achieves highly detection rate with low false alarm rate.

Ref [8] have presented a fast distributed outlier detection algorithm for mixed attribute datasets that deals with sparse high-dimensional data. The algorithm called outlier detection for mixed attribute datasets identifies outliers based on the categorical attributes first, and then focuses on subsets of data in the continuous space by utilizing information about these subsets from the categorical attribute space.

Ref [9] improved the speed of intrusion detection system, keep the high detect date and the low false positive rate using the Parallel Clustering Ensemble based on Evidence Accumulation algorithm, it overcomes the disadvantages of conventional Parallel K-means algorithm. Through paralleling, the algorithm clusters more speedily facing to mass data, and keep the advantages of the Evidence Accumulation which combines the results of multiple clustering into a single data partition.

Ref [10] present a method for outlier detection that uses HPSO clustering based on swarm intelligence, which is capable of providing clustering at different levels of compactness. Merging clusters and attribute evolution help in learning about the correct cluster solution and outlier data. Experiments show that the approach is capable of identifying true outliers as well as a good clustering configuration of data. Setting parameters automatically is a challenging problem in this method. High dimension data is also problematic in this research.

Ref [11] is an anomaly detection algorithm based on hierarchical clustering, called ADBHC. ADBHC generates clusters using density-based partitioning method which has less computational cost. It uses the improved hierarchical clustering tree to carry out fast scalable and adaptive anomaly detection. The improved hierarchical clustering tree supports updating profiles at any time. They extend the clustering algorithm and apply branch and bound mechanism for filtering noise. ADBHC had lower false alarm rate and higher detection rate. The superior performance of detection was mainly due to the high accuracy of normality profiles and the capability of filtering noise. Various parameters had pernicious impacts on the adaptive captivity of ADBHC.

- Advantages: Anomaly detection can detect novel attacks to increase the detection rate. Compared to supervised approaches, unsupervised approach breaks the dependency on attack-free training datasets. The performance of unsupervised anomaly detection approaches achieve higher detection rate over supervised approach. Also, unsupervised approach have high false positive rate over supervised approach. Using unsupervised anomaly detection techniques, however, the system can be trained with unlabeled data and is capable of detecting previously unseen attacks[12].

- Disadvantages: Obviously, not all typical behaviors are attacks or intrusion attempts. This represents one drawback of intrusion detection methods based on clustering[13].

c) *Hybrid methods*

Through analyzing the advantages and disadvantages between anomaly detection and misuse detection, a mixed intrusion detection system (IDS) model is designed. [14]First, data is examined by the misuse detection module, then abnormal data detection is examined by anomaly detection module.

Ref [1]proposed combinatorial approach for unsupervised classification of anomalous and normal activities in computer network. The proposed approach combines the two well-known machine learning methods: the k-Means clustering and the ID3 decision tree learning approaches. The k-Means method was first applied to partition the training instances into k disjoint clusters. The ID3 decision tree built on each cluster learns the subgroups within the cluster and partitions the decision space into finer classification regions; thereby improving the overall classification performance.

Ref [15]An incremental intrusion detecting model is proposed. This model integrates unsupervised Self Organizing Map and supervised Radial Basis Function to complete incremental learning. Self Organizing Map can get new type intrusion information and generate new nodes in Radial Basis Function. By this model, intrusion of unknown type can be detected online.

Fuzzy clustering algorithm is an unsupervised anomaly detection technique without training; it does not need to know the type of attack in Intrusion Detection data samples, so it can detect a variety of known and unknown characteristics of network intrusion simultaneously. This article combined QPSO with the FCM algorithm, using QPSO algorithm has better features to find the global optimal value, using particle swarm flying in the solution space search best value to replace FCM iterative process to obtain a more suitable mix of clustering algorithm[16].

In order to reduce or eliminate the noise impact on constructing the hyper plane of SVM, firstly it preprocesses the data, after that the fuzzy membership function is introduced into SVM. The fuzzy membership function acquires different values for each input data according to different effects on the classification result. Because different network protocol has different attributes, that must affect the detection effect. This paper proposes cooperative network intrusion detection Based on Fuzzy SVM. Three types of detecting agents are generated according to TCP, UDP and ICMP protocol. How to improve the accuracy of UDP detection agent in existing data set will be the major weakness[17].

V. Conclusions

In this paper we have demonstrated some difficulties in Network Intrusion Detection Systems where its log files are high scale and dimensions; consequently, new methods need to be developed for processing these huge data sources. Furthermore concept drift is nature of data in IDS and should be managed by new methods. On the other hand, efficiency in terms of accuracy is one of the most critical measurements which are mostly defined by ratio of false positive and false negative alarms. Therefore, we need to design efficient algorithms whereas scan data once and extract hidden patterns inside it. Evolving data, visiting data once, accuracy in intrusion detections and space limitations are major issues in intrusion detection systems. However, there are two main approaches for intrusion detection: firs group employs signature-based methods to identify attacks and second one refers to anomaly detection techniques but devising new framework with combining these two main approaches can overcame most drawbacks.

VI. Acknowledgement

This work was supported by grant 03-04-10-875FR from the Basic Research Program of the University Putra Malaysia.

References Références Referencias

1. Y. Yasami and S. P. Mozaffari, "A novel unsupervised classification approach for network anomaly detection by k-Means clustering and ID3 decision tree learning methods," The Journal of Supercomputing, vol. 53, pp. 231-245, 2010.
2. B. Thuraisingham, L. Khan, M. M. Masud, and K. W. Hamlen, "Data mining for security applications," 2009.
3. S. Y. Wu and E. Yen, "Data mining-based intrusion detectors," Expert Systems with Applications, vol. 36, pp. 5605-5612, 2009.
4. F. Xie and S. Bai, "Using Data Field to Analyze Network Intrusions," Information Security Practice and Experience, pp. 78-89, 2006.
5. S. Y. Jiang, X. Song, H. Wang, J. J. Han, and Q. H. Li, "A clustering-based method for unsupervised intrusion detections," Pattern Recognition Letters, vol. 27, pp. 802-810, 2006.
6. W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," 2002.
7. S. Sun and Y. Z. Wang, "A Weighted Support Vector Clustering Algorithm and its Application in Network Intrusion Detection," 2009.
8. Koufakou and M. Georgiopoulos, "A fast outlier detection strategy for distributed high-

dimensional data sets with mixed attributes," Data Mining and Knowledge Discovery, vol. 20, pp. 259-289, 2010.
9. H. Gao, D. Zhu, and X. Wang, "A Parallel Clustering Ensemble Algorithm for Intrusion Detection System," 2010.
10. S. Alam, G. Dobbie, P. Riddle, and M. A. Naeem, "A swarm intelligence based clustering approach for outlier detection," 2010.
11. H. Liang, R. Wei-wu, and R. Fei, "An Adaptive Anomaly Detection Based on Hierarchical Clustering," 2009.
12. P. Gogoi, B. Borah, and D. K. Bhattacharyya, "Anomaly Detection Analysis of Intrusion Data using Supervised & Unsupervised Approach," Journal of Convergence Information Technology, vol. 5, 2010.
13. G. Singh, F. Masseglia, C. Fiot, A. Marascu, and P. Poncelet, "Mining Common Outliers for Intrusion Detection," Advances in Knowledge Discovery and Management, pp. 217-234, 2010.
14. Zhang, G. Zhang, and S. Sun, "A Mixed Unsupervised Clustering-Based Intrusion Detection Model," 2009.
15. L. Y. Tian and W. P. Liu, "Incremental intrusion detecting method based on SOM/RBF," 2010.
16. H. Wang, Y. Zhang, and D. Li, "Network intrusion detection based on hybrid Fuzzy C-mean clustering," 2010.
17. S. Teng, H. Du, N. Wu, W. Zhang, and J. Su, "A Cooperative Network Intrusion detection Based on Fuzzy SVMs," Journal of Networks, vol. 5, pp. 475, 2010.