ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 4, April 2013

Survey paper on Data Mining techniques of Intrusion Detection

Harshna (M.Tech C.S.E)
Department Of Computer Science & Engineering of RIMT Institutions
MandiGobindgarh, Sirhind

NavneetKaur(Assistant Professor in CSE)
Department of Computer Science & Engineering of RIMT Institutions
MandiGobindgarh, Sirhind

***Abstract*—In Information Security, intrusion detection is the act of detecting actions that attempt to compromise the integrity, confidentiality, or availability of a resource. Intrusion detection does not, in general, include prevention of intrusions. This paper is concentrating on data mining techniques that are being used for such purposes. Advantages and disadvantages of these techniques have been discussed in this paper. Modern intrusion detection applications facing complex problems. These applications has to be require extensible, reliable, easy to manage, and have low maintenance cost.**

***Index Terms*— Data Mining, Intrusion Detection, Knowledge Discovery Database, Patterns**

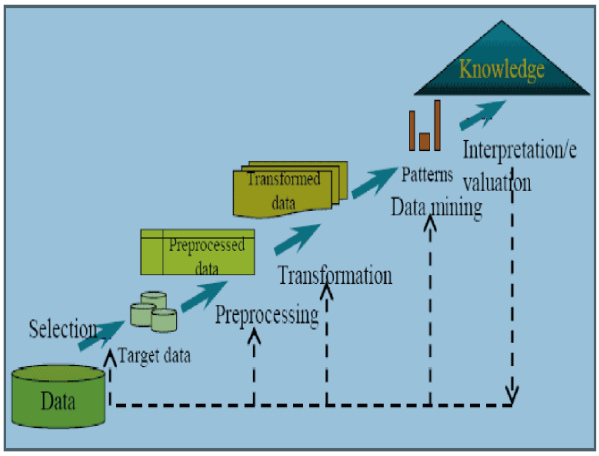
1.Introduction

Data mining has attracted a lot of attention due to increased, generation, transmission and storage of volume data and an need for extracting useful information and knowledge from them[2]. In past year's research have started looking into the possibility of using data mining techniques in the emerging field of information security especially in the challenging problem of intrusion detection. Intrusion is commonly defined as a set of actions that attempt to violate the integrity, confidentiality or availability of a system. Intrusion detection is the process of finding important events occurring in a computer system and analyzing them for possible presence of intrusion. So, it is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems.
An effective and quality based IDS needs an array of diverse components and features, including
a.Centralized view of the data
b.Data transformation capabilities
c.Analytic and data mining methods
d.Flexible detector deployment, including scheduling that enables periodic model relation and distribution
e.Real-time detection and alert infrastructure
f.Reporting capabilities
g.Distributed processing
h.High system availability
i.Scalability with system load

In general, that are two types of attacks:
(i) Inside attack are the ones in which an intruder has all the privilege to access the application or
the system, but it performmalicious actions.
(ii) Outside attack are the ones in which the intruder does not have proper rights to access the system. Detecting inside attack is usually more difficult compare to outside attack.

2. Data Mining ,KDD and related fields

Data mining (DM), also called Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using association rules . Data mining is frequently used to designate the process of extracting useful information from large databases. The term knowledge discovery in databases (KDD) is used to denote the process of extracting beneficial knowledge from large data sets. Data mining, by contrast, refers to one particular step in the process of Knowledge Discovery. Spherically, the data mining step applies so called data mining techniques to extract patterns from the data.
Here, we broadly outline some of the most basic KDD steps:
1. Understanding the application domain: First is developing an understanding of the application domain, the relevant background knowledge, and the specific goals of the KDD endeavor.
2. Data integration and selection: Second is the integration of multiple data sources and then selection of the subset of data that is relevant to the analysis task.
3. Data mining: Third is the application of particular algorithms for extracting patterns from data.
4. Pattern evaluation: Fourth is the interpretation and validation of the discovered patterns. The goal of the step is to guarantee that actual knowledge is being discovered.
5. Knowledge representation: Third step is to document the discovered knowledge.



3. Current used techniques

Traditional method for intrusion detection based on signature based method. For this extensive knowledge of signature of previously known attacks is necessary. In this monitored events are matched with the signature to detect intrusion. The feature is extract from various different audit database and then comparing these features with to a set of attack signature provide by human expert for intrusion detection.

There are other approaches on the implementation of IDS. Among those techniques, two are the most popular: *Anomaly detection* is based on the detection of traffic anomalies. The deviation of the monitored traffic from the normal profile is measured. *Misuse/Signature detection*: looks for patterns and signatures of already known attacks in the network traffic. A constantly updated database is usually used to store the signatures of known attacks. The way this technique deals with intrusion detection resembles the way that anti-virus software operates

4. Data Mining meets Intrusion Detection

4.1 Anomaly Detection or Profile Matching *:* This technique is based on the normal behaviour of a subject (e.g., a user or a system); any action that significantly deviates from the normal behaviour is considered as an intrusive action. Misuse detection catches intrusions in terms of the characteristics of known attacks or system vulnerabilities; any action that conforms to the pattern of a known attack or vulnerability is considered intrusive. The anomaly approach is focused on normal behaviours patterns. When a new kind of activity becomes acceptable (does not contradict to security policy), the normal behaviour pattern database must be updated; otherwise the activity will be treated as an intrusion and will result in false positives. Attacks and deviations from normal activity are anomaly by definition and deserve the IDS user's attention.

Although anomaly detection can find out unknown patterns of attacks, it also suffers from several drawbacks. A general problem of all anomaly detection approaches, with the exception of the specification-based technique, is that the subject's normal behaviour is modelled on the basis of the (audit) data collected over a period of normal operation. If undiscovered intrusive activities occur during this period, they will be taken as normal activities. In addition, because a subject's normal behaviour usually changes over time (for example, a user's behaviour may change when he moves from one project to another), the IDSs that use the above approach usually allow the subject's profile to gradually change. So, this gives an intruder the chance to gradually train the IDS and trick it into accepting intrusive activities as normal. Also, because these approaches are all based on summarized information, they are insensitive to stealthy attacks. Because of some technical reasons, the current anomaly detection approaches usually suffer from a high false-alarm rate. Another difficult problem in building such models is how to decide the features to be used as the input of the models (e.g., the statistical models). In the existing models, the input parameters are generally decided by domain experts (e.g., network security experts) in ad hoc ways. So, it is not guaranteed that all the features related to intrusion detection will be selected as input parameters. Missing important intrusion-related features makes it difficult to distinguish attacks from normal activities, having non-intrusion-related features could introduce "noise" into the models and thus affect the detection performance.

4.2 Misuse Detection or Signature Matching: Misuse detection is said to be complementary to anomaly detection.In misuse detection approach, firstly abnormal system behaviour is defined, and then define any other behaviour, as normal behaviour. Its main advantage is simplicity of adding known attacks to the model. Therefore, this systems look for well-defined patterns of known attacks or vulnerabilities. They can catch an intrusive activity even if it is so negligible that the anomaly detection approaches tend to ignore it. Attacks and deviations from normal behaviour are taken as anomalies.

The disadvantage of misuse detection is that it cannot detect novel or unknown attacks. As a result, the computer systems protected solely by misuse detection systems face the risk of being comprised without detecting the attacks. In addition, due to the requirement of explicit representation of attacks, the detection system requires the nature of the attacks to be well understood. It implies that human experts must work on the analysis and representation of attacks. So, it is time consuming and error prone.

Additionally, intrusion detection systems (IDSs) are categorized according to the kind of input information they analyze. This leads to the distinction between host-based and network-based IDSs. Host-based IDSs analyze host-bound audit sources such as operating system audit trails, system logs, or application logs. Network-based IDSs analyze network packets that are captured on a network.

5. Drawbacks of IDS

Intrusion Detection Systems (IDS) have become a standard component in security systems as they allow network administrators to detect policy violations and these policy violations range from external attackers trying to gain unauthorized access to insiders abusing their access. But these IDSs also have some drawbacks which are described as under as :

a. Current IDS are usually tuned to detect known service level network attacks. This leaves them vulnerable to original and unknown or novel malicious attacks.

b. Data overload: Another aspect which does not relate directly to misuse detection but is extremely important is how much data an analyst can efficiently analyze. Depending on the intrusion detection tools employed by a company and its size there is the possibility for logs to reach millions of records per day.

c. False positives: A false positive occurs when normal attack is mistakenly classified as malicious and treated accordingly.

d. False negatives: In this case, an IDS does not generate an alert when an intrusion is actually taking place. (Classification of malicious traffic as normal)

Data mining can help improve intrusion detection by addressing each and every one of the above mentioned problems.

Remove normal activity from alarm data to allow analysts to focus on real attacks

- Identify false alarm generators and "bad" sensor signatures
- Find anomalous activity that uncovers a real attack
- Identify long, on going patterns (different IP address, same activity)

To accomplish these tasks, data miners employ one or more of the following techniques:

Data summarization with statistics, including finding outliersVisualization: presenting a graphical summary of the data

Clustering of the data into natural categories

Association rule discovery: defining normal activity and enabling the discovery of anomalies
Classification: predicting the category to which a particular record belongs.

6.Survey of Applied Techniques

In this section a survey of data mining techniques that have been applied to IDSs by various research groupsis presented.

A. Feature Selection
Feature selection, also known as subset selection or variable selection. It is a process commonly used in machine learning. Feature selection is necessary because it is computationally infeasible to use all available features, or because of problems of estimation when limited data samples (but a large number of features) are present.
B. Machine Learning
Machine Learning is defined as the study of computer algorithms that improve automatically through experience. Applications varies from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests. As compared to statistical techniques, machine learning techniques are well suited to learning patterns with no a priori knowledge of what those patterns may be.Classification and Clustering are the two most popular machine learning problems.
1) Classification Techniques: In a classification task in machine learning, the task is to take each instance of a dataset and assign it to a specific class. IDS based on classification, attempts to classify all traffic as either normal or malicious. The challenge in this method is to minimize the number of false positives and false negatives.
Five general types of techniques have been tried to perform classification for intrusion detection purposes:

a) Inductive Rule Generation: The RIPPER System is probably the most popular representative of this mechanism. Lee W. et al. used this system and proposed a framework for intrusion detection using data mining techniques. It is a rule learning program, fast and is known to generate concise rule sets. One of the attractive features of this approach is that the generated rule set is easy to understand, therefore a security analyst can verify it.

b)Genetic Algorithms: Genetic algorithms were originally introduced in the field of computational biology. These algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.Since then, they have been applied in various fields with promising results. In intrusion detection, the GA is applied to derive a set of classification rules from network audit data. The support-confidence framework is utilized as a fitness function to judge the quality of each rule. Significant properties of GA are it is robust to noise, self-learning capabilities. High attack detection rate and low false-positive rate are the advantages of GA techniques. Genetic algorithm uses a string structure for representation of rules. A string representation increases the overhead of rule formation that is the overhead for more number of rules generation .Crosbie M. et al. shows genetic programming (GP) which improves the interpretability of GA by replacing the gene structures with the tree structures, which enables higher representation ability of association rules. But due to the use of the tree data structure for rule formation, reuse of many nodes is not possible. Therefore, GP is not a very efficient method for rule mining.

c) Fuzzy Logic: Fuzzy logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than precisely deduced from classical predicate logic. The application side of fuzzy set theory dealing with well thought out real world expert values for a complex problem. In Dickerson and Dickerson the authors classify the data based on various statistical metrics. They then create and apply fuzzy logic rules to these portions of data to classify them as normal or malicious. They found that the approach is particularly effective against scans and probes. An enhancement of the fuzzy data mining approach has also been applied by Florez et al. The authors use fuzzy data mining techniques to extract patterns that represent normal behaviour for intrusion detection and describe a variety of modifications that they have made to the data mining algorithms in order to increase accuracy and efficiency. Sets of fuzzy association rules are used by them that are mined from network audit data as models of "normal behaviour." Anomalous behaviour are detected by generating fuzzy association rules from new audit data and compute the similarity with sets mined from "normal" data. If the similarity values are below a threshold value, an alarm is issued. An algorithm for computing fuzzy association rules based on Borgelt's prefix trees is described to define the modifications to the computation of support and confidence of fuzzy rules, a new method for computing the similarity of two fuzzy rule sets, and feature selection and optimization with genetic algorithms.

d) Hybrid Approach
This approach is a hybrid approach which was genetic algorithm, fuzzy logic and class-association rule mining algorithm.

e) Neural Networks: The application of neural networks for IDSs has been defined by a number of researchers. Neural networks provide a solution to the problem of modelling the users' behaviour in anomaly detection as they do not require any explicit user model. Neural networks for intrusion detection were first introduced as an alternative to statistical techniques in the IDES intrusion detection expert system to model .Advanced research issues on IDSs should involve the use of pattern recognition and learning by example approaches for the following two main reasons:
• The capability of learning by example allows the system to detect new types of intrusion.
• With earning by example approaches, attack "signatures" can be extracted automatically from labelled traffic data. This basically eliminates the subjectivity and other problems
introduced by the presence of the human factor.

f) Immunological based techniques: These techniques are defined as the set of connections from normal traffic as the "self", then generate a large number of "non-self" examples: connections that are not part of the normal traffic on a machine. These examples are generated using a byte oriented hash and permutation. They can then compare incoming connections using the r-contiguous bits match rule. If a connection matches one of the examples, it is assumed to be in non-self and marked as anomalous.

g) Support Vector Machine: Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression.SVM is widely applied to the field of pattern recognition. It is also used for an intrusion detection system. "one class SVM" is based on one set of examples belonging to a particular class and no negative examples rather than using positive and negative examples. Neither of these approaches addresses the reduction of the training time of SVM, which is what prohibits real-time usage of these approaches. With regard to the training time of SVM, random sampling has been used to enhance the training of SVM.

Clustering Techniques: Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics and many more. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets , so that the data in each subset share some common trait - often proximity according to some defined distance measure. Machine learning typically regards data clustering as a form of unsupervised learning. It is useful in intrusion detection as malicious activity should cluster together, separating

itself from non-malicious activity. Clustering provides some significant advantages over the classification techniques already discussed, in that it does not require the use of a labelled data set for training. Clustering techniques are of five types: hierarchical, statistical, exemplar, distance, and conceptual clustering, each of which has different ways of determining cluster membership and representation.
Statistical Techniques: Three basic classes of statistical techniques are linear, nonlinear (such as a regression-curve), and decision trees. Statistics also includes more complicated techniques, such as Markov models and Bayes estimators. Statistical patterns can be calculated with respect to different time windows, such as day of the week, day of the month, month of the year, etc. or on a per-host, or per-service basis . Denning (1987) described how to use statistical measures to detect anomalies, as well as some of the problems and their solutions in such an approach.

Conclusion

This paper has presented a survey of the various data mining techniques like feature selection, machine learning and statistical techniques. Machine learning is further divided into two types: Classification and Clustering. In classification various techniques like inductive rule training, genetic algorithms, fuzzy logic, hybrid technique, neural networks, immunological based techniques, SVM.So, these techniques are discussedthat have been proposed towards the enhancement of IDSs. This paper presents the ways in which data mining has been known to aid the process of Intrusion Detection and the ways in which the various techniques have been applied and evaluated by researchers.

References

[1] Crosbie M. and Spafford G., **"Applying genetic programming to intrusion detection,"** presented at the AAAI Fall Symp. Series, AAAI Press, Menlo Park, CA, Tech. Rep. FS-95-01, 1995.

[2] Dickerson, J. E. and J. A. Dickerson, **"Fuzzy network profiling for intrusion detection"**, In Proc. of NAFIPS 19th International Conference of the North American Fuzzy Information Processing Society, Atlanta, pp. 301306. North American Fuzzy Information Processing Society (NAFIPS), July 2000.

[3] Bankovic Z., Stepanovic D., Bojanic S., **"Improving Network Security using Genetic Algorithm Approach"**, Computer and Electrical Engineering, pp. 438-451, 2007.

[4] Ektefa M., Memar S., **"Intrusion Detection Using Data Mining Techniques"**,IEEE Trans., 2010.

[5] Naidu N. and Dharaskar R., **"An Effective Approach to Network Intrusion Detection System using Genetic Algorithm"**, International Journal of Computer Applications (0975 - 8887) volume 1 No.2, 2010. .

[6] TheodorosLappas and KonstantinosPelechrinis, **"Data Mining Techniques for (Network) Intrusion Detection Systems"**.

[7] E.Kesavulu Reddy, Member IAENG, V.Naveen Reddy, P.GovindaRajulu, **"A Study of Intrusion Detection in Data Mining"** Proceedings of the World Congress on Engineering 2011 Vol III WCE 2011,pp London, U.K. July 6 - 8, 2011

[8] Lee W. and Stolfo S., **"Data Mining Approaches for Intrusion Detection"**,Computer Science Department Columbia University.

[9] IndrJeet Rajput, DeshdeepakShrivastava**,"Data Mining Based Database Intrusion Detection System: A Survey "**, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue4, July-August 2012, pp.1752-1755

[10] Shetty M. and Shekokar N., **"Data Mining Techniques for Real Time Intrusion Detection Systems"**, International Journal of Scientific & Engineering Research Volume 3, Issue 4, April 2012.

[11] Swati Dhopte and N.Z. Tarapore, "**Design of Intrusion Detection System using Fuzzy Class-Association Rule Mining based on Genetic Algorithm"**,International Journal of Computer Applications (0975 – 8887) Volume 53– No.14, September 2012