Intrusion Detection using Data Mining Technique

Stuti Singh, Roshan Srivastava

Abstract— In reality it is not possible to prevent security breaches completely using the existing security technologies. The intrusion detection plays an important role in network security and information system. However, many current intrusion detection systems (IDSs) are signature based systems. The signature based IDS also known as misuse detection looks for a specific signature to match, and identify an intrusion. When the signatures or patterns are provided, they can detect all known attack patterns, but there are some problems for unknown attacks. The rate of false positives is very low but these types of systems are poor at detecting new attacks, variation of known attacks or attacks that act as normal behavior. Statistical Based Intrusion detection System (SBIDS) can overcome many of the aforementioned limitations of signature based intrusion detection systems. Statistical based intrusion detection systems performs better than signature based intrusion detection system for novelty detection i.e. detection of new attack is very important for intrusion detection system. Researchers have implemented various classification algorithms for intrusion detection.

This dissertation evaluates a decision tree classifier over a benchmark dataset. It will help intrusion detection system in novelty detection i.e. detection of new attacks. KDD99 dataset is used as the training data set.

Keywords — Data Mining, Decision Tree, Intrusion Detection System, KDD99 Dataset.

I. INTRODUCTION

A. Background

The field of intrusion detection has received increasing attention in recent years. One reason is the explosive growth of the internet and the large number of networked systems that exist in all types of organizations. Intrusion detection techniques using data mining have become very popular in recent years. As intrusion detection is an important application area of data mining, they aim to meliorate the great burden of analyzing huge volumes of audit data and realizing performance optimization of detection rules. The objective of this dissertation is to try out the intrusion detection on large dataset by classification algorithms binary class support vector machine and improved its learning time and detection rate in the field of Network based IDS.

B. Basic Concepts

Following are some basic concepts on which this dissertation is based.

Types of Intrusion Detection System: Current IDSs fall into two categories.

Network Based IDS: Because they only scrutinize network traffic [1], NIDS do not benefit from running on the host. As a result, they are often run on dedicated machines that observe the network flows, sometimes in conjunction with a firewall. In this case, they are not affected by security vulnerabilities on the machines they are monitoring. Nevertheless, only a limited number of information can be inferred from data gathered on the network link. Besides, widespread adoption of end-to-end encryption further limits the amount of information that can be gathered at the network interface.

Another major shortcoming of NIDS is that they are oblivious to local root attacks. An authorized user of the system that attempts to gain additional privileges will not be deleted if attack is performed locally. An authorized user of the system may be able to set up an encrypted channel when accessing the machine remotely.

Host Based IDS: HIDS have an ideal vantage point [1]. Because an HIDS runs on the machine it monitors, it can theoretically observe and log any event occurring on the machine. However, the complexity of current operating system often makes it difficult to observe any event. There are certain difficulties faced by security tools that rely on system calls interposition to monitor a host.

Intrusion Detection Techniques: All intrusion detection system use one of the two detection techniques-

Signature/Misuse based IDS: The signature based IDS are also known as misuse detection looks for a specific signature to match and then detect the intrusion. But they are of little use for as yet unknown attack methods. Most popular intrusion detection falls in to this category. This means that an IDS using misuse detection will only detect known attacks.

Statistical/Anomaly based IDS: Another approach to intrusion detection is called anomaly detection. Anomaly detection applied to intrusion detection and computer security has been an active area of research since it was originally proposed by Denning 1987. Anomaly detection algorithm has the advantage that they can detect new types of intrusion as deviations from normal behavior. In this problem, given a set of normal data to train, and given a new set of test data to test the accuracy of system. The goal of the intrusion detection system is to determine whether the test data belong to "normal" or to an "anomalous" behavior. However anomaly detection scheme suffers from a high rate of false alarms. These false alarms occur primarily because previously unseen system behavior are also recognized as anomaly.

C. Problems with Existing Intrusion Detection System

Most of the current intrusion decision systems (IDSs) are signature-based systems.

Manuscript received on March, 2013.

Stuti Singh, Computer Science and Engineering, Phagwara, Punjab, India.

Roshan Srivastava, Computer Science and Engineering, Phagwara, Punjab,India.

The signature based IDS also known as misuse detection looks for a specific signature to match, then it detect the intrusion. When the signatures or patterns are provided, they can detect all known attack patterns, but there are some problems for unknown attacks.

The rate of false positives is very small but these types of systems are poor at detecting new attacks, variations of known attacks or attacks that can be masked as normal behavior [4]. Statistical Based Intrusion Detection Systems (SBIDS) can alleviate many of the aforementioned pitfalls of a Signature Based IDS. Statistical Based Intrusion Detection Systems rely on statistical models such as the Bayes' Theorem, decision trees, etc. to identify anomalous packets on the network.

As network attacks have increased over the past few years, intrusion detection system (IDS) is becoming a widely used measure for security of information and network. Due to large volumes of sensitive data over the network as well as complex and dynamic properties of intruders, improving the performance of IDS becomes an important open problem and need more attention from the research community.

Data mining-based intrusion detection systems (IDSs) have demonstrated high accuracy, good generalization to novel types of intrusion, good classification of normal and malicious behavior, and robust behavior in a changing environment in recent years. Still a major problem faced by them is the intensive computation required in the model generation phase.

D. Objective

The objective is to classify the information of a flow available in the form of 42 attributes (i.e. Network-based IDS) as normal or attack. The classification task requires a lot of computation in model generation due to large data size.

The time required for the model generation can be reduced by removing redundant attributes by using feature selection correlation based feature selection algorithm. This work will evaluate performance of binary/two class support vector machine and multi class SVM over the Knowledge Discovery and Data Mining 1999(KDD'99) dataset. The time required for training SVM is very large for big dataset. In training SVM there is no need of instances which are not support vectors. This work filters all such instances by using clustering. This work also evaluates performance of SVM classifier for novelty detection.

II. RELATED WORK

Intrusion detection concept was introduced by **James Anderson** in 1980[5] defined an intrusion attempt or threat to be potential possibility of a deliberate unauthorized attempt to access information, manipulate or render a system unreliable or unusable. Sights moved for using data mining in content of NIDS in the late of 1990's. Researchers suddenly recognized the need for existence of standardized dataset to train IDS tool. Minnesota Intrusion Detection System (MINDS) combines signature based tool with data mining techniques. Signature based tool (Snort) are used for misuse detection & data mining for anomaly detection.

In [6] **Jake Ryan et al** applied neural networks to detect intrusions. Neural network can be used to learn a print (user behavior) & identify each user. If it does not match then the system administrator can be alerted. A back propagation neural network called NNID was trained for this process.

Denning D.E et al [7] has developed a model for monitoring audit record for abnormal activities in the system. Sequential rules are used to capture a user's behavior [8] over time. A rule base is used to store patterns of user's activities deviates significantly from those specified in the rules. High quality sequential patterns are automatically generated using inductive generalization & lower quality patterns are eliminated. An automated strategy for generation of fuzzy rules obtained from definite rules using frequent items. The developed system [9] achieved higher precision in identifying whether the records are normal or attack one.

Dewan M et al [10] presents an alert classification to reduce false positives in IDS using improved self adaptive Bayesian algorithm (ISABA). It is applied to the security domain of anomaly based network intrusion detection.

S.Sathyabama et al [11] used clustering techniques to group user's behavior together depending on their similarity & to detect different behaviors and specified as outliers.

Amir Azimi Alasti et al [12] formalized SOM to classify IDS alerts to reduce false positive alerts. Alert filtering & cluster merging algorithms are used to improve the accuracy of the system.SOM is used to find correlations between alerts.

Alan Bivens et al [13] has developed NIDS using classifying self organizing maps for data clustering. MLP neural network is an efficient way of creating uniform input for detection when a dynamic number of inputs are present.

An ensemble approach [14] helps to indirectly combine the synergistic & complementary features of the different learning paradigms without any complex hybridization. The ensemble approach outperforms both SVMs MARs & ANNs. SVMs outperform MARs & ANN in respect of Scalability, training time, running time & prediction accuracy. This paper [15] focuses on the dimensionality reduction using feature selection. The Rough set support vector machine (RSSVM) approach deploy Johnson's & genetic algorithm of rough set theory to find the reduce sets & sent to SVM to identify any type of new behavior either normal or attack one.

Aly Ei-Senary et al [16] has used data miner to integrate Apriori & Kuok's algorithms to produce fuzzy logic rules that captures features of interest in network traffic.

Taeshik Shon et al [17] proposed an enhanced SVM approach framework for detecting & classifying the novel attacks in network traffic. The overall framework consist of an enhanced SVM- based anomaly detection engine & its supplement components such as packet profiling using SOFM, packet filtering using PTF, field selection using Genetic Algorithm & packet flow-based data preprocessing. SOFM clustering was used for normal profiling. The SVM approach provides false positive rate similar to that of real NIDSs. In this paper [18] genetic algorithm can be effectively used for formulation of decision rules in intrusion detection through the attacks which are more common can be detected more accurately.

Oswais.S et al [18] proposed genetic algorithm to tune the membership function which has been used by IDS. A survey was performed using approaches based on IDS, and on implementing of Gas on IDS.

Norouzian M.R et al [19] **defined** Multi- Layer Perceptron (MLP) for implementing & designing the system to detect the attacks & classifying them in six groups with two hidden layers of neurons in the neural networks.

Host based intrusion detection is used to trace system calls. This system does not exactly need to know the program codes of each process. Normal & malicious behavior are collected through system call & analysis is done through data mining & fuzzy technique. The clustering and genetic optimizing steps [20] were used to detect the intrude action with high detection rate & low false alarm rate.

III. PROPOSED WORK

Classification is a form of data analysis that extracts models describing important data classes. These models also called as classifiers are used to predict categorical (discrete, unordered) class labels. This analysis can help us for better understanding of large data. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, credit risk and medical diagnosis [34].Data Classification is a two-step process. They are: Learning Step and Classification Step.

Learning Step: In this step classification model is constructed. A classifier is built describing a predetermined set of data classes or concepts. In learning step or training phase, where classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels.

This step is also known as supervised learning as the class label of each training tuple is provided. This learning of the classifier is "supervised" by telling to which class each training tuple belongs. In unsupervised learning or clustering, the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance.[34,35]

Classification Step: In this step, the model is used to predict class labels for given data and hence it is used for classification. First, the predictive accuracy of the classifier is estimated. To measure the classifier's accuracy, if we use the training set it would be optimistic, because the classifier tends to over fit the data i.e., during learning it may incorporate some particular anomalies of the training data that are not present in the general data set. Therefore, a test set is used, made up of the test tuples and their associated class labels. They are independent of the training tuples, from which the classifier cannot be constructed. The accuracy of a classifier on a given test set is the percentage of test tuples that are correctly classified by the classifier. The class label of each test tuple is compared with the learned classifier's class prediction for the tuple. If the accuracy of the model or classifier is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known[34,36].

Decision Tree Induction: A decision tree is a flow-chart-like tree structure, where topmost node is the root node, each internal node denotes a test to be done on an attribute, each branch represents an outcome of the test on the attribute, and leaf nodes represent classes or class distributions. Flow chart used for intrusion detection in this paper is given below:

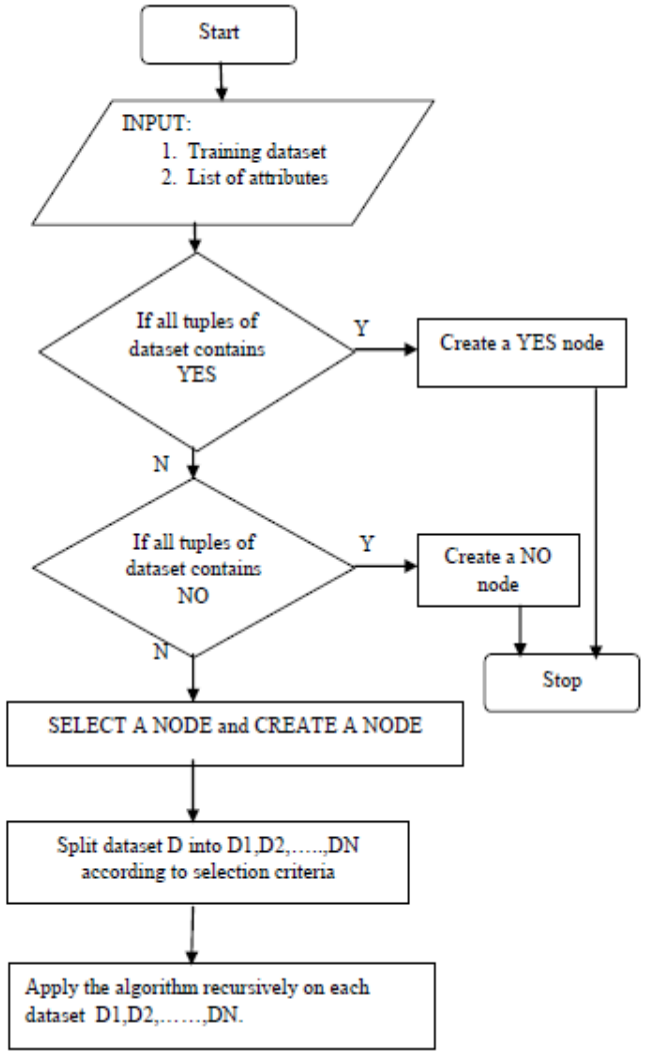

Fig. 1 Flow Chart of Proposed Method

REFERENCES

1. Litty Lionel, "Hypervisor-based Intrusion Detectio", Master of Science Graduate department of computer Science University of Torronto, 2005.
2. Mark Crosbie and gene Spafford, "Active defence of a computer system using anonymous agents", Technical report 95-008,COAST Group, Department of Computer Science, Purdue University, West Lafayette, Indiana, February 1995.
3. Litty, Intrusion Detection Http://www.cs.torronto.edu/~litty/papers/MS.pdf.
4. Network Security by Christos Douligeris, Dimitrios Nikolaou Serpanos page 93.
5. Anderson.J.P, "Computer Security Threat Monitoring & Surveilance", Technical Report, James P Anderson co., Fort Washington, Pennsylvania, 1980.
6. Jake Ryan, Meng - Jang Lin, Risto Miikkulainen, "Intrusion Processing Detection With Neural Networks", Advances in Neural Information System 10, Cambridge, MA:MIT Press,1998,DOI:10.1.1.31.3570.
7. Denning .D.E, "An Intrusion Detection Model", Transactions on Software Engineering, IEEE Communication Magazine, 1987,SE-13, PP-222-232,DOI:10.1109/TSE.1987.232894.
8. Teng.H.S, Chen.K and Lu.S.C, "Adaptive Real-Time Anomaly Detection using Inductively Generated Sequential Patterns, in the Proceedings of Symposium on research in Computer Security & Privacy, IEEE Communication Magazine,1990, pp-278-284.
9. Sekeh.M.A,Bin Maarof.M.A, "Fuzzy Intrusion Detection

System Via Data Mining with Sequence of System Calls", in the Proceedings of International Conference on Information Assurance & security (IAS)2009,IEEE Communication Magazine, pp-154-158,ISBN:978-0-7695-3744-3,DOI:10.1109/IAS.2009.32.
10. Dewan Md, Farid, Mohammed Zahidur Rahman, "Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm", Journal of Computers, Vol 5, pp-23-31, Jan 2010, DOI:10.4.304/jcp 5.1.
11. Sathyabama.S, Irfan Ahmed.M.S, Saravanan.A,"Network Intrusion Detection Using Clustering: A Data Mining Approach", International Journal of Computer Application (0975-8887), Sep-2011, Vol: 30, No: 4, ISBN: 978-93-80864-87-5, DOI: 10.5120/3670-5071.
12. Amir Azimi, Alasti, Ahrabi, Ahmad Habibizad Navin, Hadi Bahrbegi, "A New System for Clustering & Classification of Intrusion Detection System Alerts Using SOM", International Journal of Computer Science & Security, Vol: 4, Issue: 6, pp-589-597, 2011.
13. Alan Bivens, Chandrika Palagiri, Rasheda Smith, Boleslaw Szymanski, "Network-Based Intrusion Detection Using Neural Networks", in Proceedings of the Intelligent Engineering Systems Through Artificial Neural Networks, St.Louis, ANNIE-2002, and Vol: 12, pp- 579-584, ASME Press, New York.
14. Srinivas Mukkamala, Andrew H. Sung, Ajith Abraham, "Intrusion Detection Using an Ensemble of Intelligent Paradigms",Journal of Network & Computer Applications ,pp-1-15, 2004.
15. Shilendra Kumar, Shrivastava ,Preeti Jain, "Effective Anomaly Based Intrusion Detection Using Rough Set Theory & Support Vector Machine(0975-8887), Vol:18,No:3, March 2011,DOI: 10.5120/2261-2906.
16. Aly Ei-Semary, Janica Edmonds, Jesus Gonzalez-Pino, Mauricio Papa, "Applying Data Mining of Fuzzy Association Rules to Network Intrusion Detection", in the Proceedings of Workshop on Information Assurance United States Military Academy 2006, IEEE Communication Magazine, West Point, NY,DOI:10.1109/IAW.2006/652083.
17. Taeshik Shon, Jong Sub Moon, "A Hybrid Machine Learning Approach to Network Anomaly Detection", Information Sciences 2007, Vol: 177, Issue: 18, Publisher: USENIX Association, pp- 3799-3821, ISSN:00200255,DOI:10.1016/j.ins-2007.03.025.
18. Sadiq Ali Khan, "Rule-Based Network Intrusion Detection Using Genetic Algorithm", International Journal of Computer Applications, No: 8, Article: 6, 2011, DOI: 10.5120/2303-2914.
19. Norouzian.M.R, Merati.S, "Classifying Attacks in a Network Intrusion Detection System Based on Artificial Neural Networks", in the Proceedings of 13th International Conference on Advanced Communication Technology(ICACT), 2011,ISBN: 978-1-4244-8830-8,pp-868-873.
20. Jin-Ling Zhao, Jiu-fen Zhao ,Jian-Jun Li, "Intrusion Detection Based on Clustering Genetic Algorithm", in Proceedings of International Conference on Machine Learning & Cybernetics (ICML),2005, IEEE Communication Magazine,ISBN:0-7803-9091-1,DOI: 10.1109/ ICML.2005.1527621.

AUTHORS PROFILE

Stuti Singh is from Kanpur, she has done her B.Tech in Computer Science and Engineering from UPTU and pursuing M.Tech from Lovely Professional University, Jalandhar. Her research are includes Network Security and Data Mining.

Roshan Srivastava is from Kanpur. He has done his B.Tech in Computer Science and Engineering from UPTU and MS in Cyber Law and Information Security from IIIT, Allahabad and working as Assistant Professor in Lovely Professional University. His research area includes Information Security and Cryptography.