Intrusion Detection System using Classification Technique

Rajesh Wankhede
P.G. Student
Dept. Of CSE
G. H. Raisoni Academy of Engineering and Technology, Nagpur, India

Vikrant Chole
Assistant Professor
Dept. of CSE
G. H. Raisoni Academy of Engineering and Technology, Nagpur, India

ABSTRACT
In today's world people are extensively using internet and thus are also vulnerable to its flaws. Cyber security is the main area where these flaws are exploited. Intrusion is one way to exploit the internet for search of valuable information that may cause devastating damage, which can be personal or on a large scale. Thus Intrusion detection systems are placed for timely detection of such intrusion and alert the user about the same. Intrusion Detection using hybrid classification technique consist of a hybrid model i.e. misuse detection model (AdTree based) and Anomaly model (svm based).NSL-KDD intrusion detection dataset plays a vital role in calibrating intrusion detection system and is extensively used by the researchers working in the field of intrusion detection. This paper presents Association rule mining technique for IDS.

General Terms
Intrusion Detection System, Preprocessing, Security et. al.

Keywords
SVM, AdTree, NSL-KDD

1. INTRODUCTION
IDSs may be a piece of hardware or software systems which is used to detect intruders on the network. IDS systems can be distinguished according to where they're installed i.e either on the host or on the network, as well as they differ on how they detect intruders i.e. misuse detection and anomaly detection. While different types of IDS systems exist, each type of IDS has its own benefits and drawbacks. There are two types of IDS i.e Host based IDS and Network based IDS.A host based Intrusion Detection System is a system that monitors a system that it is installed on to detect the misuse or intrusion by notifying the authority or by logging the activity. One can think of a Host based IDS as mediary that monitors whether anything or anyone, whether domestic or foreign, has bypassed the system's security policy. A Host based IDS analyses the traffic directed towards and traffic sent from the specific computer on which the IDS is installed. A host-based system also has the capability to oversee key system files and any attempt to overwrite these files. A Network based IDS consist of hardware sensors located at discrete points on the network, while it may also contain the software that is installed on various computers connected along the network. These types of IDS analyses the data packets both entering and leaving the system and offering real time detection.

Intrusion detection system is mainly of two types firstly there is Anomaly detection based and second one is the Signature based systems. The Signature based Intrusion Detection System monitors packets in the network and compares them to the known signatures which are pre-configured and pre-identified based on attack behavior of previously known attacks. On the other hand the anomaly based IDS monitors the normal network traffic such as bandwidth range, types of protocols, ports and devices used to connect and sends an alert to the system administrator on detection of anomalous behavior. The signature based IDS detects attacks on the known attack signature type. Advantage of this type of system is that it can detect known attacks with low error rate, but it cannot detect the newly created attacks that do not have similar behavior to known attacks. In contrast Anomaly based IDS can be useful in identifying the new attack pattern, but in this case the error rate is higher. Thus in order to solve the above two limitations we are building a hybrid intrusion detection method that combines misuse detection method and anomaly detection method has been proposed.

Machine learning is the capability of a machine that automatically improves its performance all the way through learning from experience. In general, supervised machine learning methods are initially trained to create rules and patterns that capture characteristics of the training set. The rules and patterns are useful to identifying intrusions in test data. Data mining[9] techniques such as decision trees [13], genetic fuzzy rules [11], neural networks [12], support vector machine, principal component analysis [10], naïve Bayesian classifiers and many other feature reduction [14] algorithms have been used widely to determine the network logs and to catch intrusion related information to get better correctness of IDS. The signature based IDS detects attacks on the known attack signature type. Advantage of this type of system is that it can detect known attacks with low error rate, but it cannot detect the newly created attacks that do not have similar behaviour to known attacks. In contrast Anomaly based IDS can be useful in identifying the new attack pattern, but in this case the error rate is higher. Thus in order to solve the above two limitations we are building a hybrid intrusion detection method that combines misuse detection method and anomaly detection method has been proposed. In this paper we are going to pre-process NSL- KDD dataset to generate patterns for IDS.

2. RELATED WORK
Gisung Kim et.al, [1] proposed a new hybrid method that hierarchically combines a misuse detection and anomaly detection in a decomposed structure. First, the C4.5 decision tree was used to create the misuse detection model that is used to disintegrate the normal training data into smaller subsets. Then, the one-class support vector machine was used to create an anomaly detection model in each decomposed region. Throughout the integration, the anomaly detection model can indirectly use the known attack information to enhance its ability when building profiles of normal behavior. This is the first attempt to use the misuse detection model to enhance the ability of anomaly detection model.C4.5 decision tree does not form a cluster, which can degrade the profiling ability thus abbreviating the efficiency of the system.

Shi-Jinn Horng et.al, [2] proposed an intrusion detection system, which combines a clustering algorithm, a simple feature selection algorithm, and the Support Vector Machine (SVM). In this study, in addition to a simple feature selection method, they proposed an SVM-based network intrusion detection system with BIRCH hierarchical clustering for data pre-processing. The BIRCH hierarchical clustering provides a highly qualified and reduced datasets, in place of original large dataset, for SVM training. In addition to reduction of the training time, the resultant classifiers showed better performance than the SVM classifiers using the originally redundant dataset. However, in terms of accuracy, the proposed system could obtain the best performance at 95.72%. This approach provides better performance in terms of accuracy in comparison to the other NIDS (Network based IDS). It only detects Dos and Probe attacks not U2L and R2L attacks.

Mrutyunjaya Panda et.al, [3] proposed hybrid intelligent decision technologies using data filtering by adding guided learning methods along with a classifier to make more classified decisions in order to detect network attacks. It is seen from the results obtained that the Naive Bayes model is quite appealing because of its integrity, elegance, robustness and effectiveness. On the other hand, decision trees have proven their efficiency in both generalization and detection of new attacks. The results show that there is no single best algorithm to outperform others in all situations. In certain cases there might be dependence on the characteristics of the data. To choose a suitable algorithm, a domain expert or expert system may employ the results of the classification in order to make better decisions.

Juan Wang et.al, [4] presented an intrusion detection system based on decision tree technology. In the process of constructing intrusion rules, information gain ratio is used in place of information gain. The experiment results show that the C4.5 decision tree is feasible and effective, and has a high accuracy rate. His experimental study shows that the C4.5 decision tree is an effective technique for the implementation of decision tree and it gives almost 90% of classifier accuracy. But in this approach the error rate remains the same.

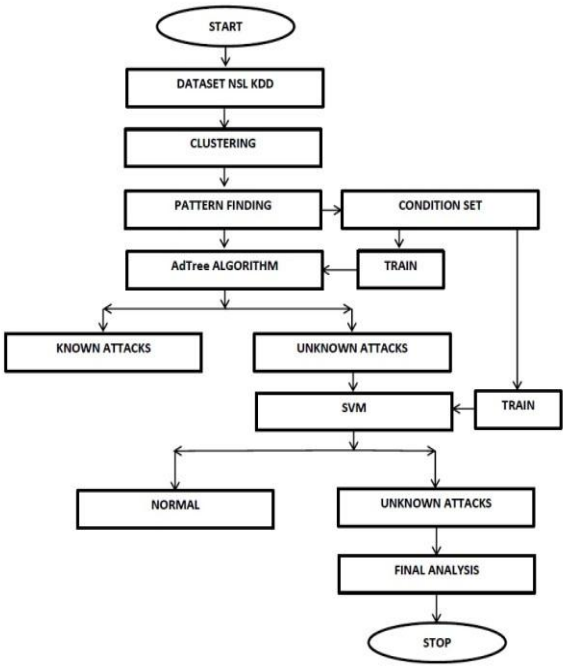
Hong Kuan Sok et.al,[5] presents a paper on using the ADTree algorithm for feature reduction. ADTree also gives good classification performance. In addition, its comprehensible decision rules endows the user to discover the features that heads towards better classification. This knowledge base facilitates to design a smaller dimension of support vectors for suitable classifier. The experiment supports the idea of using this algorithm as both knowledge discovery tool and classification. The classification task has been simplified and the speed increased drastically due to the reduced operations required to implement the classification.

Tavallaee et.al, [6] presented a paper on KDD CUP 99 Data Set and after the analysis of the entire KDD dataset it showed that there were two important issues in the data set which affected the performance of evaluated systems, and thus results in a very poor interpretation of anomaly detection approaches. To overcome the issues, NSL-KDD was proposed, which contains selected records of the KDD data set. Although, the proposed data set suffers from some problems and may not be a ideal representative of existing networks, due to the lack of public data sets for network-based IDSs, they believe that the dataset still can be used as an effective benchmark to help analyst analyze different intrusion detection methods.

Yonav Freund et.al, [8] proposes an alternating decision tree with boosting. The new learning algorithm combines boosting and decision trees. In their paper they compared the alternating decision tree with the C5.0 algorithm. On smaller datasets ADtree quickly fits the data and ADtree reaches a very small error after 50 iterations while the error of the stump boost remains large even after 200 iterations. This is a case in which large capacity of ADtree gives it an advantage. Comparing to the size of classifiers in all but three cases the classifiers generated by the ADtree are much smaller than those generated by C5.0 by boosting. The error performance of this algorithm is close to that of C5.0 with boosting.

3. PROPOSED SYSTEM

The proposed system uses Microsoft Visual Studio 2012 as front end and SQL Server Management Studio 2012 as backend for maintaining database.



Flowchart of the proposed system

Dataset: NSL-KDD Database: NSL-KDD is a dataset proposed by Tavallaee et al. NSL-KDD dataset is a reduced version of the original KDD 99 dataset. NSL-KDD consists of the same features as KDD 99. The KDD99 dataset consists of 41 features and one class tribute. The class attribute has 21 classes that fall under four types of attacks: Probe attacks, User to Root (U2R) attacks, Remote to Local (R2L) attacks and Denial of Service (DoS) attacks. This dataset has a binary class attribute. Also, it has a reasonable number of training and test instances which makes it practical to run the experiments on.

The NSL-KDD has the following differences over the original KDD 99 dataset.

1. It does not contain unwanted records in the training set; therefore the classifiers are not biased towards the frequent records.
2. No duplicate records exist in the proposed test sets; therefore, the performances of the learners are not biased towards the methods which have improved detection rates on the frequent records.

3. The number of selected records from each difficulty level group is inversely proportional to the percentage of records in the original KDD99 data set.

The sum of records in the train and test sets are acceptable, which makes it efficient to run the experiments on the complete set without the need to haphazardly select a small portion. This dataset can be obtained from the archives and is readily available on Wayback machines server.

Clustering: Cluster is an assembly of objects that belongs to the same class. In other words, identical objects are grouped in one cluster and divergent or dissimilar objects are grouped in another cluster. Clustering is a procedure of making a group of abstract objects into classes of similar objects. After gathering the dataset, it was imported to sql database. For this SQL Server Management Studio 2012 was used. For training KDDTrain+_20Percent was used. Dataset gathered was a supervised dataset; therefore the clustering was done by selecting the particular attack and maintaining the table for the same. To do this SQL query used was: "insert into [dbo].[Attack_back] select * from MainTbl where attack_type='back'" for attack back. Similar queries were used for making the clusters of other attacks. This greatly reduced the time for forming the cluster compared to using some algorithm for clustering. This query selects all the attacks labeled "back" from the main database and store it into table named [dbo].[Attack_back].

Cluster	Frequent Patern	srno	AttackName	RecordCount
View	View	1	normal	13449
View	View	2	neptune	8282
View	View	3	ipsweep	710
View	View	4	satan	691
View	View	5	portsweep	587
View	View	6	smurf	529
View	View	7	nmap	301
View	View	8	back	196
View	View	9	teardrop	188
View	View	10	warezclient	181

Screenshot: Clustered data and its count

Pre-processing: Data pre-processing is a method of processing which is performed on raw data to prepare it for another processing operation. Commonly used as a preparatory mining practice, data pre-processing converts the data into a format that will be more easily and adequately processed as per the user requirements. The pre-processing of the data is done using association rule mining. Association rules are if/then statements that helps discover the correlations between seemingly irrelevant data in a relational database or other information archive. An association rule has two parts, an anterior (if) and a subsequent (then). An anterior part is an item found in the data. A subsequent is an item that is found in combination with the anterior part. Association rules are constructed by analyzing data for frequent if/then patterns and using the criteria support and confidence to classify the most important relationships. Support is an expression of how frequently the items appear in the database. Confidence hints the number of times the if/then statements have been found to be true.

Frequent Patterns Of Attacksmurf

srno	duration	protocoltype	service	flag	srcbytes	dstbytes	land	wrongfragment	urgent	hot	numfailed_logins	loggedin	numcompromised	rootshell	suattempted	numroot	numfile_creations	numshells	numaccess_files	numoutbound_cmds	ishost_l
10	0	icmp	ecr_i	SF	520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	icmp	ecr_i	SF	520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	icmp	ecr_i	SF	520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	icmp	ecr_i	SF	520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	icmp	ecr_i	SF	1032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	icmp	ecr_i	SF	1032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	icmp	ecr_i	SF	1032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	icmp	ecr_i	SF	1032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Screenshot: Frequent patterns of attack Smurf

Association rule mining is generally used to find the interesting rules from a large database depending upon the user defined support and confidence. In market basket analysis it finds relationship among the items present in the transactional database. A frequent item set is defined as one that occurs more frequently in the given data set than the user given support value. One more threshold confidence is used to restrict the association rules to a limited number. Confidence also includes the item sets having low support but from which high confidence rules may be generated.

4. CONCLUSION

NSL-KDD dataset was taken and preprocessed, first the clustering was done and then association mining was applied in order to generate the frequent patterns for various known attacks. These frequent patterns provide a baseline for preventing the attacks from entering the system and also distinguish attacks from normal data, thus allowing normal data to enter the system. The future work will be done on hybrid classification technique for intrusion detection system which will include the misuse detection model based on Alternating Decision tree and anomaly detection model based on Support Vector Machine.

5. REFERENCES

[1] Gisung Kim and Seungmin Lee (2014), A Novel Hybrid Intrusion Detection Method Integrating Anomaly Detection With Misuse Detection, ELSEVIER, Expert Systems with Applications vol. 41 pp. 1690 – 1700.

[2] Shi-Jinn Horng and Ming-Yang Su (2011), "Novel Intrusion Detection System Based On Hierarchical Clustering and Support Vector Machines", ELSEVIER, Expert Systems with Applications. pp. 38 306-313.

[3] Mrutyunjaya Panda and Manas Ranjan Patra, "A Comparative Study Of Data Mining Algorithms For Network Intrusion Detection", First International Conference on Emerging Trends in Engineering and Technology, pp 504-507, IEEE, 2008.

[4] Juan Wang, Qiren Yang, Dasen Ren, "An intrusion detection algorithm based on decision tree technology", In the Proc. of IEEE Asia-Pacific Conference on Information Processing, 2009.

[5] Hong Kuan Sok et.al, "Using the ADTree for Feature Reduction through Knowledge Discovery" Instrumentation and Measurement Technology Conference (I2MTC), 2013 IEEE International ,pp1040 – 1044.

[6] Tavallaee M, Bagheri E, Lu W, Ghorbani A. "A detailed analysis of the KDD CUP 99 data set", 2009 IEEE Symposium on Computational intelligence for security and defense applications, 2009,pp 1-6.

[7] F. Amiri, M. Yousefi, C. Lucas, A. Shakery and N. Yazdani, "Mutual Information-Based Feature Selection for Intrusion Detection Systems", Journal of Network and Computer Applications, Vol. 34, 2011, pp.1184–1199.

[8] Yonav Freund et.al, "The Alternating Decision Tree Algorithm", ICML '99 Proceedings of the Sixteenth International Conference on Machine Learning, pp 124-133.

[9] Hachmi Fatma, Limam Mohamed "A two-stage technique to improve intrusion detection systems based on data mining algorithms" IEEE, 2013. pp 1-6.

[10] H.F. Eid, A. Darwish A. H. Ella and A. Abraham, "Principle components analysis and Support Vector Machine based Intrusion Detection System," 2010, 10th International Conference on Intelligent Systems Design and Applications (ISDA), 2010.

[11] Tsang, C. H., Kwong, S., & Wang, H.," Genetic-fuzzy rule reordering in mining approach and evaluation of feature selection techniques for anomaly intrusion detection", Pattern Recognition,40 (9), pp. 2373–2391, 2007.

[12] Zhi-Song Pan, Song-Can Chen, Gen-Bao Hu, DaoQiang Zhang, (2010), "Hybrid Neural Network and C4.5 for Misuse Detection ", Proceedings of the second International conference on Machine Learning and Cybernetics, November, pp. 2463 – 2467.

[13] Siva S.Sivatha Sindhu, S.Geetha and A. Kannan," Decision tree based light weight intrusion detection using a wrapper approach", Expert Systems with Applications, 39(1), pp.129-141, 2012.

[14] Therese Bjerkestrand et,al. "An Evaluation of Feature Selection and Reduction Algorithms for Network IDS Data" 2015 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA 2015).