New Scheme to Identify Intrusion Outliers by Machine Learning Technique

M. Thangamani, Ph.D
Assistant Professor
Department of Computer Science
Kongu Engineering College
Perundurai,
Erode-638 052
Tamilnadu, India

E.T.Venkatesh, Ph.D
Assistant Professor(SRG)
Department of Computer Science
Kongu Engineering College
Perundurai,
Erode-638 052
Tamilnadu, India

A.Kalyana Saravanan
Assistant Professor
Department of Computer Science
Kongu Engineering College
Perundurai,
Erode-638 052
Tamilnadu, India

ABSTRACT
In network administration, computers and network systems need to be protected against malicious attacks. The success of an intrusion detection system depends on the selection of the appropriate features in detecting the intrusion activity. The selection of unnecessary features may cause computational issues and reduce the accuracy of detection. In the existing work, a novel detection approach is used through a one-class learning algorithm based on support vector machine classification. It can be also used in a Bayesian framework to estimate the posterior class probabilities of test data with unknown class. This algorithm can detect the system anomalies and monitor the health of a system. It does not allow updating the training data with new information. Therefore, the accuracy of the algorithm is low for the test data. The proposed work aims to improve the performance of attack detection and to reduce the false-alarm rate using hybrid classifier. This approach effectively identifies the set of attacks such as Denial of Service, Probe, and User to Root and Remote to Local attacks. In addition, an Experimental evaluation is carried out to compare the performance of existing classifier with the proposed Decision tree-Bayesian network classifier.

1. INTRODUCTION
Network security is a more specialized area that consists of policies and regulations adopted by the network administrator to monitor and prevent unauthorized access. The purpose of network security is to protect our information's from unauthorized access.. Due to enormous growth in network usage and the huge increase in the number of applications running on top, security is very critical and also essential. So that the systems, suffer from security vulnerabilities which are both technically difficult and economically expensive.

Arun Hodigere [1] suggested that intrusion is an attempt that break into our system without proper authorization and misuse the information. The Intruder may be from outside the network or inside. It can be a physical or remote intrusion. There are different ways to intrude such as buffer overflows, unexpected combinations, and unhandled input and race conditions. Intrusion Detection (ID) is the process of detecting security breaches by examining events occurring in a computer system. An ID system gathers and analyzes information from various areas within a system or network to identify possible security breaches. They include both intrusion and misuse. It uses vulnerability assessment, which is a technology developed to assess the security of a computer or network.

Some of the important features an intrusion detection system should possess include those that are fault tolerant and they run continually with minimal human supervision. The IDS must be able to recover from system crashes, either accidental or caused by malicious activity. They need to possess the ability to resist subversion so that an attacker cannot disable or modify the IDS easily. Furthermore, the IDS must be able to detect any modifications forced on the IDS by an attacker. They are minimal overhead on the system to avoid interfering with the normal operation of the system. They are also configurable so as to accurately implement the security policies of the systems that are being monitored. The IDS must be adaptable to changes in the system and user behavior over time. They are easy to deploy this can be achieved through portability to different architectures and operating systems, through simple installation mechanisms, and by being easy to use by the operator. General enough to detect different types of attacks and must not recognize any legitimate activity as an attack (false positives). At the same time, the IDS must not fail to recognize any real attacks (false negatives). It stops the attack, whether it is a program or a hacker. It closes the loop-hole through which the attacker gained access. Either pop-up to an online admin, or email or SMS to a remote admin. What is done by the attack to the network, and from where the attack came, helps gather forensic evidence should a prosecution become necessary or possible.

2. RELATED WORK
Christoper et.al [2] presented a method for performing Bayesian classification of input events for intrusion detection. It reports many false alerts as the traditional approach. A variety of intrusion detection system (IDS) [3] has been employed for protecting computers and networks from malicious and host based attacks .Gonzalez and Dasgupta [4] suggested that an anomaly detection problem can be stated as a two-class problem in which the given data can be classified as normal or attack. Classification of the anomaly detection techniques [5] according to the nature of the processing involved in the ''behavioral'' model considered for targeted system. Data mining framework for constructing intrusion detection models are outlined in [6]. They have applied data mining programs to audit data to compute misuse and anomaly detection models, according to the observed behavior in the data. James P.Anderson [7] proposed an computer security threat monitoring and surveillance from the various summary accounting and audit trail reports. Then developed a new approach to classify network requests using support vector machine by [9, 10].

Marsland [11] posted that the novelty detection is concerned with recognizing inputs that differ in some way from those that are usually seen. It serves a useful technique in cases where an important class of data is under represented in the training set. This means that the performance of the network will be poor for those classes. If the training data only consist of examples from one class and the test data contain examples from two or more classes, the classification task is called anomaly detection or novelty detection. Fault detection methods, such as built-in tests, typically log the time at which the error occurred and they either trigger alarms for manual intervention or initiate automatic recovery. With enterprise networks, network analyzers are often attached to the lines in order to monitor traffic and send an alarm when disruptions are detected. An hybrid architecture developed [12] for combining different feature selection algorithms for real world. Vasilis et. al [14] investigates the use of a one-class support vector machine algorithm to detect the onset of system anomalies, and trend output classification probabilities, as a way to monitor the health of a system. Data mining IDS framework [8, 13, 15, 16, 17] have been presented.

Outlier detection is an important activity in many critical and safely environments [18], where the outlier indicates the abnormal running conditions from which a major performance deprivation may result. An outlier is an ID that denotes an anomaly node in a network and also the anomaly data in distributed systems. An outlier quickly pinpoints an intruder inside a system with malicious intentions. Derisstiawan [19] presented a comprehensive review of the early detection, response and prevention system feature. However, this system has no benchmarking with regard to real-network traffic. Ganapathy et. al [20], proposed an intelligent agent-based weighted distance outlier-detection algorithm. While Thangamani et al. [21, 22, 23, 24, 25] mooted a machine learning algorithm for semantic representation in a peer to peer environment.

3. METHODOLOGY

The enormous growth in networks over the computer system provides the opportunities for intrusions and other malicious activities. Intrusion detection is the important task in the system security. Therefore, the methods based on hand-coded rule sets are laborious to build and are not very reliable. This problem has led to an increasing interest in intrusion detection techniques based on machine learning or data mining. However, traditional data mining based intrusion detection systems used single classifier in their detection engines. So the authors propose a hybrid classifier for intrusion detection relying on Bayesian Network and C4.5 Decision Tree.

3.1 Decision tree

Decision tree induction is one of the classification algorithms in data mining. This classification algorithm is inductively learned to construct a model from the pre-classified data set. Each data item is defined by values of the attributes. Classification may be viewed as mapping from a set of attributes to a particular class. The decision tree classifies the given data item using the values of its attributes. The decision tree is initially constructed from a set of pre-classified data. This main approach selects the attributes, which best divide the data items into their classes. According to the values of these attributes, the data items are partitioned. This process is recursively applied to each partitioned subset of the data items. The process terminates when all the data items in the current subset belong to the same class.

A node of a decision tree specifies an attribute by which the data are to be partitioned. Each node has a number of edges, which are labeled according to a possible value of the attribute in the parent node. An edge connects either two nodes or a node and a leaf. Leaves are labeled with a decision value for categorization of the data. Induction of the decision tree uses the training data, which are described in terms of the attributes. The main problem here is to decide the attribute, which best partitions the data into various classes. The ID3 algorithm [Quinlan 1986] uses the information theoretic approach to solve this problem. Information theory exploits the concept of entropy, which measures the impurity of a data item. The value of entropy is small when the class distribution is uneven when all the data items belong to one class. The entropy value is higher when the class distribution is more even, that is when the data items have more classes.

Information gain is a measure of the utility of each attribute in classifying the data items. It is measured using the entropy value. Information gain measures the decrease of the weighted average impurity (entropy) of the attributes compared with the impurity of the complete set of data items.

$$\mathrm{Gain(A)} = \mathrm{Info(D)} - \mathrm{InfoA\ (D)} \quad (1)$$

$$\mathrm{Info(D)} = -\sum_{i=1} p_{ij} \log 2(p_i) \quad (2)$$

Therefore, the attributes with the largest information gain are considered as the most useful for classifying the data items. To classify an unknown object, one starts at the root of the decision tree and follows the branch indicated by the outcome of each test until a leaf node is reached. The name of the class at the leaf node is the resulting classification. Decision tree induction has been implemented with several algorithms. The attribute with the highest normalized information gain is chosen to make the decision. In this algorithm all the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the indicating the need to choose that class. The general algorithm for building a decision tree is to check for base cases. For each attribute A, find the normalized information gain by splitting A. Let A_best be the attribute with the highest normalized information gain. Create a decision node that splits on A_best. Recur on the sub lists obtained by splitting on a A_best, and add those nodes as the children of the node.

3.1 C4.5 Decision Tree

Its an algorithm used to generate a decision tree developed by Ross Quinlan. It is a successor of ID3 algorithm (Iterative Dichotomiser). C4.5 can be used to generate a decision tree, and for this reason it is also called statistical classifier. C4.5 builds decision tree from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S=s1,s2,…. of already classified samples. Each sample si=x1,x2,…. Represents the attributes of each sample. The training data are augmented with a vector C=c1, c2,…, where c1,c2,… represents the class. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits the set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain that results from choosing an attribute for splitting the data. Of particular interest to this work is the C4.5 decision tree algorithm. C4.5 avoids over fitting the data by determining a decision tree. It handles continuous attributes and is able to choose an appropriate attribute selection measure. It handles training data with missing attribute values and improves computation efficiency. C4.5

builds the tree from a set of data items using the best attribute to test and divide the data item into subsets. Then it uses the same procedure on each sub set recursively. The best attribute that divides the subset at each stage is selected using the information gain of the attributes.

3.1 Decision tree in Intrusion Detection

Intrusion detection can be considered as a classification problem where each connection or user is identified either as one of the attack types or as normal based on some existing data. Decision trees can solve this classification problem of intrusion detection as they learn the model from the data set. It can classify the new data item into one of the classes specified in the data set. Decision trees can be used as misused intrusion detection as they can learn a model based on the training data. They can predict the future data as one of the attack types or as normal ones based on the learned model.

A decision tree works well with large data sets. This is important as a large amount of data flow across computer networks. The high performance of decision tree makes it useful in real-time intrusion detection. Decision trees construct easily interpretable models, which are useful for a security officer to inspect and edit. These models can also be used in the rule-based models with minimum processing. The generalization accuracy of decision trees is yet another useful property of intrusion detection model. There will always be some new attacks on the system which are small variations of known attacks after the intrusion detection models are built. The ability to detect these new intrusions is possible due to the generalization accuracy of decision trees.

4. EXPERIMENTAL RESULT AND DISCUSSION

The evaluation of IDS with regard to the attacks they detect is a first and vital step in the context of IDS. Weka-3.7 [15] data mining tool kit is used for analyzing the results. The correctly and incorrectly classified instances show the percentage of trained and test instances that were correctly and incorrectly classified. The percentage of correctly classified instances is often called accuracy.

4.1 Data set Description

The NSL-KDD of a new version KDD'99 data set is used. It solves some of the inherent problems of the KDD'99 data set. It can be applied to an effective data set to help researchers compare different intrusion detection methods. It runs the experiments on the complete set without the need to randomly select a small portion. The training and testing set has 41 attributes.

These attributes are divided into three main groups, namely Intrinsic features, Content features and Traffic features. Intrinsic attributes encapsulate all the attributes that can be extracted from a TCP/IP Connection. In content attribute, category attributes are extracted from the content area of the network packets based on expert domain knowledge .Unlike most of the DoS and Probing attacks, the R2L and U2R attacks don't have any intrusion frequent sequential patterns. This is because the DoS and Probing attacks involve many connections to some host(s) in a very short period of time; however the R2L and U2R attacks are embedded in the data portions of the packets, and normally involves only a single connection. Traffic features attributes are calculated by taking into account the previous connections and it is divided into time traffic features and machine traffic features.

4.2 Experimental Result

The experimental results obtained when directly applying the two methods (existing and proposed system) on the testing dataset, made up 11,850 records. Here, the only interest lies in knowing to which category (normal, DoS, R2L, U2R, Probe) a given connection belongs. The accuracy of each experiment is based on the percentage of successful classification (PSC) on the test data set where

$$\text{PSC} = \frac{\text{Number of correctly classified Instances}}{\text{Number of instances in the test set}} \qquad (3)$$

Figure 1 shows the attack detection rate calculated in SVM classification and hybrid classification methods. The attack detection rate for the proposed system is 55.40% and for existing SVM classification it is only 49.73%. When comparing these results, hybrid classifier produces a good detection rate.

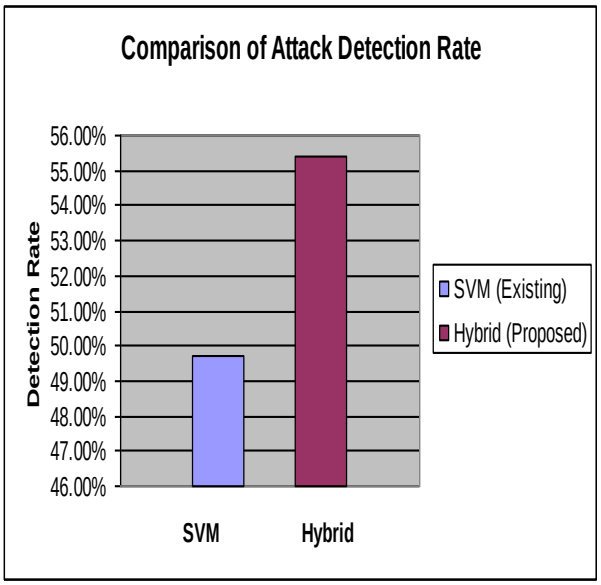

Figure 1 Comparison of Attack Detection Rate

5. CONCLUSION AND FUTURE WORK

This work dealt with a hybrid classification approach for anomaly based intrusion detection system. In the existing system, the support vector machine and Bayesian classification are used to detect intrusions. This algorithm is useful for anomaly detection in the absence of failure information and it can be used in real time. It can be used on any multivariate training and test data with a minimum input from the positive class training data. Therefore, the accuracy is not up to the expectation. The proposed work an improved version of the Bayesian model used in the existing intrusion detection system by utilizing Support Vector Machine (SVM). It effectively improves the detection rate. First, the system is trained with trained dataset which detects the known attack types. Then, it is tested using test dataset which detects the unknown attack types. This meta classifier detects the DoS, Probe, R2L and U2R attacks from both trained data and test data.

Future research, investigate other data mining techniques with a view to enhance the detection accuracy as close as possible

to 100% while maintaining a low false positive rate. It will also explore problem domains in the real world.

6. REFERENCES

[1] ArunHodigere et al (2001),"Intrusion Detection System",shiraz university.

[2] Christopher Kruegel,DarrenMutz, William Robertson, and FrerickValeur(2003), "Bayesian Event Classification for Intrusion Detection", Proceedings of 19th Anniversary of Computer Security Applications Conference,pp.14-23.

[3] Dorothy E. Denning, and P.G.Neumann (1985), "Requirement and model for IDES- A real time intrusion detection system," Computer Science Laboratory, SRI International, Menlo Park, Technical Report No. 83F83-01-00.

[4] F.Gonzalez and D.Dasgupta (2003),"Anomaly Detection using real-valued negative selection",Genetic Programming and Evolvable Machines, Volume 4, pp-383-403.

[5] Garci'a-TeodoroP,Di'az-Verdejo J, and Macis'-Ferna'ndez G(2009), "Anomaly-based network intrusion detection: Techniques,Systems,and Challenges", International journals on Computers and Security, Vol.28,pp.18-28.

[6] HuyAnhNguyen,Deokjai Choi, "Application of Data Mining to Network Intrusion Detection: Classifier Selection Model".

[7] James P.Anderson (1980),"Computer security threat monitoring and surveillance," Technical Report 98-17,Fort Washington, Pennsylvania, USA.

[8] Jiawei Han and MichelineKamber, "Data Mining: Concepts and Techniques", 2nd Edition,University of IIIinois.

[9] John Mill,Atsushi Inoue(2006), "Support Vector Classifiers and Network Intrusion Detection", Eastern Washington University press.

[10] J.Kwok(2006), "Moderating the outputs of support vector machine classifiers", IEEE Transactions on Neural networks, Volume 10, pp. 1018–1031.

[11] S.Marsland (2003), "Novelty detection in learning systems ", Neural computing surveys, Volume 3, pp.157-195.

[12] SrilathaChebrolu (2005), 'Feature deduction and ensemble design of intrusion detection systems', Elsevier Journal of Computers & Security Vol. 24/4, pp. 295-307.

[13] Sudipto Banerjee and AnrdrewO.Finley (2009), " Bayesian Linear Models", University of Minnesota,U.S.A.

[14] Vasilis A. Sotiris, Peter W.Tse, and Michael G.Pecht (2010), "Anomaly Detection Through a Bayesian Support Vector Machine," IEEE Transactions on Reliability, volume 59, no.2, pp. 277–286.

[15] WEKA: Data Mining Software in Java (2008), ttp://www.cs.waikato.ac.nz/ml/weak

[16] Wenke Lee, S.Stolfo, and K.Mok (1999) , "A Data Mining Framework for Building intrusion Detection Model ", Proceedings of IEEE symposium, Security and Privacy, pp 120-132.

[17] WillmottSteveng, Ackleso .R, Davis, David R.Legate and Clinton M.Rowe(1985), "Statistics for the Evaluation and Comparison of models," journal of Geophysical Research, vol. 90, no. C5, pages 8995-9005.

[18] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-based detection and prediction of outliers," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 2, pp. 145–160, 2006.

[19] Derisstiawan, ala'yaseenibrahimshakhatreh (2012) , " intrusion prevention system: a survey", Journal of Theoretical and Applied Information Technology, issue:1, Vol. 40, Pp. 44-54.

[20] Ganapathy .S, Yogesh. P and Kannan .A (2012), "Intelligent Agent-Based Intrusion Detection System Using Enhanced Multiclass SVM", Computational Intelligence and Neuroscience, Hindawi Publishing Corporation, pp.1-10.

[21] Thangamani .M and Thangaraj .P,"Survey on Text Document Clustering",International Journal of Computer Science and Information Security,vol.8(4),2010.

[22] Thangamani, M. and Thangaraj, P. "Integrated Clustering and Feature Selection Scheme for Text Documents", International Journal of Computer Science, Vol.6, Issue 5, pp.536-541, 2010.

[23] Thangamani.M and Thangaraj.P, "Effective fuzzy semantic clustering scheme for decentralized network through multidomain ontology model", International Journal of Metadata, Semantics and Ontologies, Interscience Vol.7, Issue 2, pp.131-139, December 2012 Interscience publication

[24] Thangamani.M and Thangarj.P. "Fuzzy ontology for document clustering based on genetic Algorithm", International Journal of Applied mathematics and information science, Vol.4, Issue 7, pp.1563-1574, 2013.

[25] Thangamani.M and Thangarj.P. "Effective Fuzzy Ontology for Distributed Document Using Non-Dominated Ranked Genetic Algorithm", International Journal of Intelligent Information Technologies (JIIT), Vol.7 (4), pp.26-46, 2011