Data Mining Techniques for Real Time Intrusion Detection Systems

Monali Shetty, Prof. N.M.Shekokar

Abstarct- Due to the widespread proliferation of computer networks, attacks on computer systems are increasing day by day. Preventive measures can stop these attacks to some extent, but they are not very effective due to various reasons. This lead to the development of intrusion detection as a second line of defense. In Information Security, intrusion detection is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource. Intrusion detection does not, in general, include prevention of intrusions. In this paper, we are focused on data mining techniques that are being used for such purposes. We debate on the advantages and disadvantages of these techniques. Finally we present a new idea on how data mining can aid IDSs in real time.

In this paper, we present an overview of real time data mining-based intrusion detection system (IDSs). We focus on issues related to deploying a data mining -based IDS in a real time environment. New intelligent Intrusion Detection Systems (IDSs) are based on sophisticated algorithms rather than current signature-base detections are in demand. In this paper, we propose a new real time data-mining based technique for intrusion detection using an ensemble of binary classifiers with feature selection and multiboosting simultaneously.

Index Terms - Data Mining , DOS attack, Feature Selection ,Intrusion Detection Systems, Multiboosting, Network Security, Real time IDS

1. INTRODUCTION

A secure network must provide the following:

• Data confidentiality: Data that are being transferred through the network should be accessible only to those that have been properly authorized.

• Data integrity: Data should maintain their integrity from the moment they are transmitted to the moment they are actually received . No corruption or data loss is accepted either from random events or malicious activity.

• Data availability: The network should be resilient to Denial of Service attacks.

An intrusion detection system (IDS) inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system.

The rest of this work is a survey of data mining techniques that have been applied to IDSs and is organized as follows: In section 2 we debate on the drawbacks of standard IDSs. Section 3 offers a brief introduction to data mining and section 4 illustrates how data mining can be used to enhance IDSs. In section 5 we talk about the various data mining techniques that have been employed in IDSs by various researchers. Section 6 presents existing IDSs that use data mining techniques. In section 7, we give new proposal on how data mining can be used to aid IDSs, while in section 8 we conclude our work.

2. DRAWBACKS OF IDSS

Intrusion Detection Systems (IDS) have become a standard component in security infrastructures as they allow network administrators to detect policy violations. These policy violations range from external attackers trying to gain unauthorized access to insiders abusing their access.

Current IDS have a number of significant drawbacks:

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

• 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. That amount of data he needs to look at seems to be growing rapidly. 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.

• False positives: A common complaint is the amount of false positives an IDS will generate. A false positive occurs when normal attack is mistakenly classified as malicious and treated accordingly.

• False negatives: This is the case where 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.

3. WHAT IS DATA MINING?

Data Mining, is the process of automatically searching large volumes of data for patterns using association rules.

Here are a few specific things that data mining might contribute to an intrusion detection project:

- Find anomalous activity that uncovers a real attack
- Remove normal activity from alarm data to allow analysts to focus on real attacks
- Identify long, ongoing 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 outliers
- Visualization: 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

4. DATA MINING AND IDS

The main function of the data mining model that we are interested in is classification, as normal, or malicious, or as a particular type of attack [10][12].
Common representations for data mining techniques include rules, decision trees, linear and non-linear functions (including neural nets), instance-based examples, and probability models [10].

4.1. Off Line Processing

The use of data mining techniques in IDSs, usually implies analysis of the collected data in an offline environment. There are important advantages in performing intrusion detection in an offline environment, in addition to the real-time detection tasks typically employed.
Below we present the most important of these advantages:

- In off-line analysis, it is assumed that all connections have already finished and, therefore, we can compute all the features and check the detection rules one by one [13].
- The estimation and detection process is generally very demanding and, therefore, the problem cannot be addressed in an online environment because of the various the real time constraints [16][15]. Many real-time IDSs will start to drop packets when flooded with data faster than they can process it.
- An offline environment provides the ability to transfer logs from remote sites to a central site for analysis during off peak times.

4.2. Data Mining and Real Time IDSs

Even though offline processing has a number of significant advantages, data mining techniques can also be used to enhance IDSs in real time. Lee et al. [17] were one of the first to address important and challenging issues of accuracy, efficiency, and usability of real-time IDSs. They implemented feature extraction and construction algorithms for labeled audit data.
.e.,g entropy, conditional entropy, relative entropy, information gain, and information cost to capture intrinsic characteristics of normal data and use such measures to guide the process of building and evaluating anomaly detection models. A serious limitation of their approaches (as well as with most existing IDSs) is that they only do intrusion detection at the network or system level.
However, with the rapid growth of e-Commerce and e-Government applications, there is an urgent need to do intrusion detection at the application-level. This is because many attacks may focus on applications that have no effect on the underlying network or system activities.

4.3. Sensor Correlation

The use of multiple sensors to collect data by various sources has been presented by numerous researchers as a way to increase the performance of an IDS.

- Kumar [12] states that ,"Correlation of information from different sources has allowed additional information to be inferred that may be difficult to obtain directly."
- Lee et al.[17], state that using multiple sensors for ID should increase the accuracy of IDSs.

4.4. Evaluation Datasets

To test the effectiveness of data mining techniques in IDSs the use of established and appropriate datasets is required.
.

1. The DARPA datasets, available from the Lincoln Laboratory at MIT (http://www.ll.mit.edu/IST/ideval), are the most popular and widely used.
2 .'Knowledge Development and Data mining' (KDD) '99 cup challenge dataset.

These evaluations are contributing significantly to the intrusion detection research field by providing direction for research.

5. SURVEY OF APPLIED TECHNIQUES

In this section we present a survey of data mining techniques that have been applied to IDSs by various research groups.

5.1. Machine Learning

Machine Learning is the study of computer algorithms that improve automatically through experience. Applications range from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests. In contrast to statistical techniques, machine learning techniques are well suited to Clustering and Classification are probably the two most popular machine learning problems. Techniques that address both of these problems have been applied to IDSs.

5.1.1 Classification Techniques

In a classification task in machine learning, the task is to take each instance of a dataset and assign it to a particular class. A classification based IDS attempts to classify all traffic as either normal or malicious. The challenge in this is to minimize the number of false positives (classification of normal traffic as malicious) and false negatives (classification of malicious traffic as normal).
Five general categories of techniques have been tried to perform classification for intrusion detection purposes:

a) Neural Networks : The application of neural networks for IDSs has been investigated by a number of researchers. Neural networks provide a solution to the problem of modeling the users' behavior in anomaly detection because 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 [29]. Numerous projects have used neural nets for intrusion detection using data from individual hosts [10].

McHugh et al [30] have pointed out that advanced research issues on IDSs should involve the use of pattern recognition and learning by example approaches for one reason:

• The capability of learning by example allows the system to detect new types of intrusion.

A different approach to anomaly detection based on neural networks is proposed by Lee et al. While previous works have addressed the anomaly detection problem by analyzing the audit records produced by the operating system, in this approach, anomalies are detected by looking at the usage of network protocols.

b) Fuzzy Logic : Fuzzy logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than precisely deduced from classical predicate logic.

An enhancement of the fuzzy data mining approach has also been applied by Florez et al. [27] The authors use fuzzy data mining techniques to extract patterns that represent normal behavior for intrusion detection.

Luo [28] also attempted classification of the data using Fuzzy logic rules.

c) Genetic Algorithm : **Genetic algorithms were originally introduced in the field of computational biology. Since then, they have been applied in various fields with** promising results. Fairly recently, researchers have tried to integrate these algorithms with IDSs.

Chittur [25] applied a genetic algorithm and used a decision tree to represent the data. They used the "Detection rate minus the false positive rate" as their preference criterion to distinguish among the data.

The REGAL System [23][24] is a concept learning.

d) Support Vector Machine : Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. SVMs attempt to separate data into multiple classes.

Mukkamala, Sung, et al. [32] used a more conventional SVM approach. They used five SVMs, one to identify normal traffic, and one to identify each of the four types of malicious activity in the KDD Cup dataset.

Eskin et al. [31], and Honig et al. [19] used an SVM in addition to their clustering methods for unsupervised learning. The achieved performance was comparable to or better than both of their clustering methods.

5.1.2 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. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Machine learning typically regards data clustering as a form of unsupervised learning. Clustering 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 labeled data set for training.

Eskin et al. [31], and Chan et al. have applied fixed width and k-nearest neighbor clustering techniques to connection logs looking for outliers, which represent anomalies in the network traffic.

Bloedorn et al. [15] use a similar approach utilizing k-means clustering.

Marin et al. [34] employed a hybrid approach that begins with the application of expert rules to reduce the dimensionality of the data, followed by an initial clustering of the data and subsequent refinement of the cluster locations using a competitive network called Learning Vector Quantization. Since Learning Vector Quantization is a nearest neighbor classifier, they classified a new record presented to the network that lies outside a specified distance as a masquerader. Thus, this system does not require anomalous records to be included in the training set.

The authors were able to achieve classification rates, in some cases near 80% with misclassification rates less than 20%.

5.2. Feature Selection

"Feature selection, also known as subset selection or variable selection, is a process commonly used in machine learning, wherein a subset of the features available from the data is selected for application of a learning algorithm. Feature selection is necessary either 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. Table I contains some examples of the features selected. Each of these features offers a valuable piece of information to the System.

TABLE 1.

FEATURES THAT HAVE BEEN EXTRACTED IN THE PROCESS OF APPLYING DATA MINING TECHNIQUES TO IDSS

Destination IP	# ICMP packets

Destination port	# to certain services
bytes transferred / all services	# total connections
Protocol	wrong data packet size rate
Source bytes	# urgent
TCP Flags	% data packet
Source port	# to privileged services
Duplicate ACK rate	# other errors
bytes transferred / current host	# packets to all services
% of same service to same host	# different services accessed
average duration / all services	# FIN flags

5.3. Statistical Techniques

Statistical techniques, also known as "top-down" learning ,are employed when we have some idea as to the relationship were looking for and can employ mathematics to aid our search.

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. [33], or on a per-host, or per-service basis [13].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. The five statistical measures that she described were the operational model, the mean and standard deviation model, the multivariate model, the Markov process model, and the time series model.

Sinclair et al. [64] describe how they used Quinlan's ID3 algorithm to build a decision tree to classify network connection data. Bloedorn et al.[15] and Barbara et al. [65] also use decision tree-based methods.

1) Hidden Markov Models: Much work has been done or proposed involving Markovian models. For instance, the generalized Markov chain may improve the accuracy of detecting statistical anomalies. Unfortunately, it has been noted that these are complex and time consuming to construct [11], however their use may be more feasible in a high-power off-line environment. A hidden Markov model (HMM) is a statistical model where the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications. A HMM can be considered as the simplest dynamic Bayesian network.

5.4. Ensemble Approaches

"In reality there are many different types of intrusions, and different detectors are needed to detect them." [Axelsson]

One way to improve certain properties, such as accuracy, of a data mining system is to use a multiplicity of techniques and correlate the results together. The combined use of numerous data mining methods is known as an ensemble approach, and the process of learning the correlation between these ensemble techniques is known by names such as learning, or meta-learning**.**

Lee and Stolfo [13][14][24] state that if one method or technique fails to detect an attack, then another should detect it. They propose the use of a mechanism that consists of multiple classifiers, in order to improve the effectiveness of the IDS.

6. EXISTING SYSTEMS

In this section, we present some of the implemented systems that apply data mining techniques in the field of Intrusion Detection.

1. The MINDS System [36]: The Minnesota Intrusion Detection System (MINDS), uses data mining techniques to automatically detect attacks against computer networks and systems. While the long-term objective of MINDS is to address all aspects of intrusion detection, the system currently focuses on two specific issues:

2. EMERALD (SRI) [35]: EMERALD is a software-based solution that utilizes lightweight sensors distributed over a network or series of networks for real-time detection of anomalous or suspicious activity. EMERALD sensors monitor activity both on host servers and network traffic streams. By using highly distributed surveillance and response monitors, EMERALD provides a wide range of information security coverage, real-time monitoring and response, protection of informational assets.

3. IDSs in the Open Market: Various systems that employ data mining techniques have already been released as parts of commercial security packages.– Dshield,, RealSecure SiteProtectort

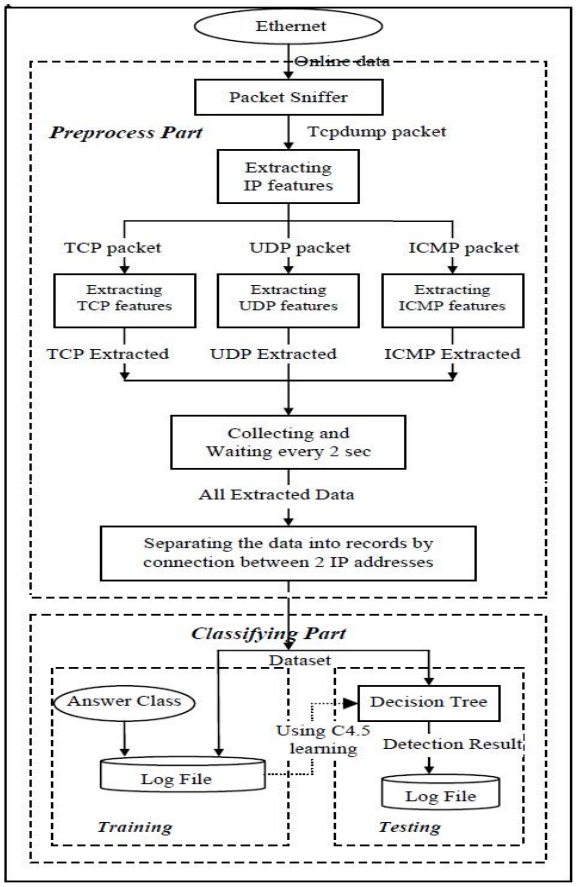

Fig. 2 Real-time IDS process(Existing System)

7. PROPOSED MODEL

In this section we propose a data mining technique that could potentially prove to be beneficial to Real Time IDSs. The idea is to use a new data-mining based technique for intrusion detection using an ensemble of binary classifiers with feature selection and multiboosting simultaneously.
We are making changes in Classifying Part .Our model employs feature selection so that the binary classifier for each type of attack can be more accurate, which improves the detection of attacks that occur less frequently in the training data. Based on the accurate binary classifiers, our model applies a new ensemble approach which aggregates each binary classifier's decisions for the same input and decides which class is most suitable for a given input. During this process, the potential bias of certain binary classifier could be alleviated by other binary classifiers' decision. Our model also makes use of multiboosting for reducing both variance and bias. In this model ,
For each trial i, i=1…T, where T is the total no. of trials,
(1) A sample training set is generated by a multibooster using wagging (as specified in Webb's multiboosting algorithm [15]).
(2) Binary classifiers are generated for each class of event using relevant features for the class and the C4.5 classification algorithm [13].

Binary classifiers are derived from the training sample by considering all classes other than the current class as other, e.g., Cnormal will consider two classes: normal and other. The purpose of this phase is to select different features for different classes by applying the information gain [18] or gain ratio [13] in order to identify relevant features for each binary classifier. Moreover, applying the information gain or gain ratio will return all the features that contain more information for separating the current class from all other classes. The output of this ensemble of binary classifiers will be decided using arbitration function based on the confidence level of the output of individual binary classifiers (e.g., see Fig. 2).
(3) The ensemble classifier is used by the multibooster in order to calculate the classification error, and derive the next training set.
(4) After T trials, the final committee is formed and it will be used by our intrusion detection system.

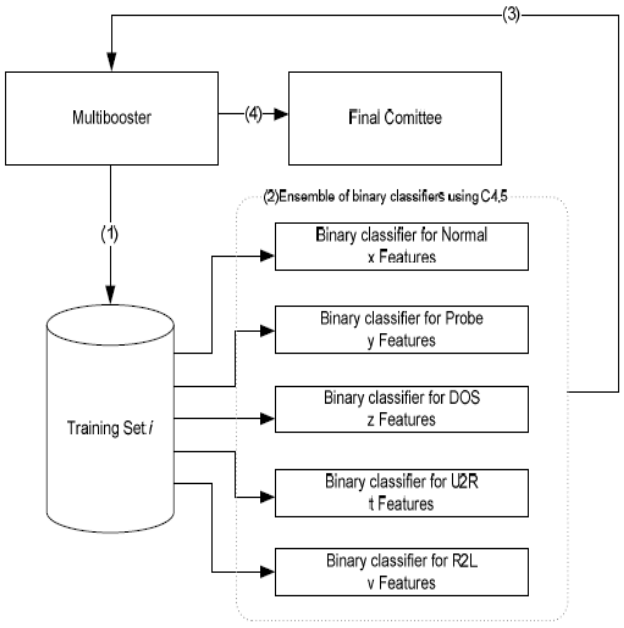

Fig. 3 The Diagram Of The Proposed Model

8. CONCLUSIONS

This paper has presented a survey of the various data mining techniques that have been proposed towards the enhancement of IDSs. We have shown the ways in which data mining has been known to implement the process of Intrusion Detection and the ways in which the various techniques have been applied and evaluated by researchers. Finally, in the last section, we proposed a data mining approach that we feel can contribute significantly in the attempt to create better and more effective Intrusion Detection Systems.

9. REFERENCES

[1] Christos Douligeris, Aikaterini Mitrokotsa, "DDoS attacks and defense mechanisms: classification and state- of-the-art",ComputerNetworks:The International Journal of Computer and Telecommunications Networking,Vol. 44, Issue 5, pp: 643 - 666,2004

[2] Denning, D. E., "An intrusion-detection model", IEEE Transactions on Software Engineering 13 (2), 222-232, February, 1987.

[3] http://www.securityfocus.com/news/2445

[4] http://www.webopedia.com

[5] Mithcel ,Rowton, Introduction to Network Security Intrusion Detection , December 2005

[6] Biswanath Mukherjee, L.Todd Heberlein, Karl.Levitt, "Network Intrusion Detection",IEEE, June 1994.

[7] Presentation on Intrusion Detection Systems, Arian Mavriqi

[8] Intrusion Detection Methodologies Demystfied Enterasys, Networks TM.

[9] Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," Communications of the ACM 39 (11),November 1996, 2734

[10] Ghosh,A.K.,A.Schwartzbard, and M.and Schatz,"Learning program behavior profiles for intrusion detection ", In Proc. 1st USENIX, 9-12 April, 1999.

[11] Kumar, S., "Classification and Detection of Computer Intrusion", PhD. thesis, 1995, Purdue Univ., West Lafayette, IN.

[12] Lee, W. and S. J. Stolfo, "Data mining approaches for intrusion detection", In Proc. of the 7th USENIX

Security Symp., San Antonio, TX. USENIX, 1998.

[13] W. Lee, S.J.Stolfo et al, "A data mining and CIDF based approach for detecting novel and distributed intrusions", Third International Workshop on Recent Advancesin Intrusion Detection (RAID 2000),

[14] Lee, W., S. J. Stolfo, and K. W. Mok, "A data mining

framework for building intrusion detection models," In

1999 IEEE Symp. On Security and Privacy, Oakland,

CA, pp. 120132. IEEE Computer Society Press, 9-12

May 1999

[15] Eric Bloedorn et al, "Data Mining for Network Intrusion Detection: How to Get Started," Technical paper, 2001.

[16] Singh, S. and S. Kandula, "Argus - a distributed network-intrusion detection system," Undergraduate Thesis, Indian Institute of Technology, May 2001.

[17]Lee, W. and D. Xiang, "Information-theoretic measures for anomaly detection",2001 IEEE Symp. on Security and Privacy, IEEE 2001

[18] Dickerson, J. E. and J. A. Dickerson, "Fuzzy network profiling for intrusion detection",NAFIPS 19th International Conference of the North American Fuzzy Information Processing Society, Atlanta, pp. 301306.

[19] Honig, A., A. Howard, E. Eskin, and S. J. Stolfo, "Adaptive model generation: An architecture for the deployment of data mining based intrusion detection systems", In D. Barbar and S. Jajodia (Eds.), Data Mining for Security Applications.May 2002.

[20] Helmer, G., J.Wong, V. Honavar, and L. Miller, "Automated discovery of concise predictive rules for intrusion detection", Technical Report 99-01, Iowa State Univ., Ames, IA, January, 1999.

[21] Cohen, W. W., "Fast effective rule induction", Proc. of the 12th International Conference on Machine Learning, 9-12 July, 1995.

[22] S. Stolfo, A. L. Prodromidis and P. K. Chan, "JAM: Java Agents for Meta-Learning over Distributed Databases", Third International Conference on Knowledge Discovery and Data Mining, 1997

[23] Neri, F., "Comparing local search with respect to genetic evolution to detect intrusion in computer networks", In Proc. of the 2000 Congress on Evolutionary Computation CEC00, IEEE Press,16-19 July, 2000.

[24] Neri, F., "Mining TCP/IP traffic for network intrusion detection", Machine Learning: ECML 2000.

[25] Chittur, A.,"Model generation for an intrusion detection system using genetic algorithms", High School Honors Thesis, Ossining High School. In cooperation with Columbia Univ, 2001.

[26] Theodoros Lappas and Konstantinos Pelechrinis, "Data Mining Techniques for Intrusion Detection Systems"

[27] G. Florez, SM. Bridges, Vaughn RB, "An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection", Annual Meeting of The North American Fuzzy Information Processing Society Proceedings, 2002.

[28] Luo, J., "Integrating fuzzy logic with data mining methods for intrusion detection", Masters thesis, Mississippi State Univ., 1999.

[29] Debar, H., Becker, M., and Siboni, D., "A Neural Network Component for an Intrusion Detection System", IEEE Computer Society Symposium on Research in Security and Privacy, Los Alamitos, CA, pp. 240–250, Oakland,CA, May 1992.

[30] McHugh J., "Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory", ACM Trans. Information System Security 3 (4), 262294, 2000.

[31] Eskin, E., A. Arnold, M. Preraua, L. Portnoy, and S. J. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data", In D. Barbar and S. Jajodia (Eds.), Data Mining for Security Applications. Boston: Kluwer Academic Publishers, May 2002.

[32] Mukkamala, S., A. H. Sung, "Identifying key variables for intrusion detection using neural networks",15th International Conference on Computer Communications, pp. 1132-1138, 2002.

[33] Frank, J., "Artificial intelligence and intrusion detection: Current and future directions",17th National Computer Security Conference, Baltimore, MD. (NIST), 1994.

[34] Marin, J. A., D. Ragsdale, and J. Surdu, "A hybrid approach to profile creation and intrusion detection", DARPA Information Survivability Conference and Exposition, Anaheim, CA. IEEE Computer Society,12-14 June, 2001.

[35] Porras, A. and Neumann, P. G., "EMERALD: Event Monitoring Enabling Responses to Anomalous Live Disturbances",National Information Systems Security Conference, October 1997.

[36] Levent Ertoz and Eric Eilertson and Aleksandar Lazarevic and Pang-Ning Tan and Vipin Kumar and Jaideep Srivastava and

Paul Dokas, "MINDS - Minnesota Intrusion Detection System",
Next Generation Data Mining,MIT Press, 2004.