A Novel Datamining Based Approach for Remote Intrusion Detection

Renu Deepti.S, Loshma.G
CSE , Sri Vasavi Engineering College(JNTUK)
Tadepalligudem-India

***Abstract*—Today, as information systems are more open to the Internet,attacks and intrusions are also increasing rapidly so the importance of secure networks is also vital. New intelligent Intrusion Detection Systems which are based on sophisticated algorithms are in demand.Intrusion Detection System (IDS) is an important detection used as a countermeasure to preserve data integrity and system availability from attacks. It is a combination of software and hardware that attempts to perform intrusion detection.In data mining based intrusion detection system, we should make use of particular domain knowledge in relation to intrusion detection in order to efficiently extract relative rules from large amounts of records.This paper proposes boosting method for intrusion detection and it is possible to detect the intrusions in all the Systems, without installing the Software in client System (like client-server) via Web service (Apache tomcat) by using the ip address of the client system.**

Keywords—boosting, data mining, anomaly detection, network intrusion detection system

I. INTRODUCTION

The network technologies have provided us with new life and shopping experiences, particularly in the fields of e-business, e-learning and e-money. But along with network development,there has come a huge increase in network crime. It not only greatly affects our everyday life, which relies heavily on networks and Internet technologies, but also damages computer systems that serve our daily activities, including business, learning, entertainment and so on.

The intrusion detection techniques based upon data mining [1, 2] are generally fall into one of two categories: misuse detection and anomaly detection. Misuse detection is based on extensive knowledge of patterns associated with known attacks provided by human experts. Pattern matching, data mining, and state transition analysis are some of the approaches for Misuse detection. Anomaly detection is based on profiles that represent normal behavior of users, hosts, networks, and detecting attacks of significant deviation from these profiles. Intrusion Detection Systems (IDS) is a combination of software and hardware that attempts to perform intrusion detection.

Unlike signature-based intrusion detection systems, models of misuse are created automatically, and can be more sophisticated and precise in case of datamining techniques. The main motivation behind using intrusion detection in data mining [5, 10, 12, 13, 15] is automation..To apply data mining techniques in intrusion detection, first, the collected monitoring data needs to be preprocessed and converted to the format suitable for mining processing. Next, the reformatted data will be used to develop a clustering or classification model. The classification model can be rule-based, decision-tree based, association-rule based, Bayesian-network based, or neural network based.Intrusion Detection mechanism based on IDS are not only automated but also provides for a significantly elevated accuracy. Data mining techniques can be applied to gain insightful knowledge of intrusion prevention mechanisms. They can help detect new vulnerabilities and intrusions, discover previous unknown patterns of attacker behaviors, and provide decision support for intrusion management..

This paper introduces the remote intrusion detection[2] by using which we can detect the intrusions present in one system any other system without installing this software in the other system.

II. DATA MINING?

Data Mining, is one of the hot topic in the field of knowledge extraction from database. Data mining is used to automatically learn patterns from large quantities of data. Mining can efficiently discover useful and interesting rules from large collection of data.

Data mining is disciplines works to finds the major relations between collections of data and enables to discover a new and anomalies behavior. Data mining based intrusion detection techniques generally fall into one of two categories; misuse detection and anomaly detection. In misuse detection, each instance in a data set is labeled as ‘normal’ or ‘intrusion’ and a learning algorithm is trained over the labeled data. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks, as long as they have been labeled appropriately. Data mining are used in different field such as marketing, financial affairs and business organizations in general and proof it is success. The main approaches of data mining that are used including classification which maps a data item into one of several predefined categories. This approach normally output “classifiers” has ability to classify new data in the future, for example, in the form of decision trees or rules. An ideal application in intrusion detection will be together sufficient “normal” and “abnormal” audit data for a user or a program. The second important approach is Clustering which maps data items into groups according to similarity or distance between them.

Data mining techniques can be differentiated by their different model functions and representation, preference criterion,and algorithms [10].The main function of the model that we are interested in is classification, as normal, or malicious, or as a particular type ofattack [11][12].We are also interested in link and sequence analysis [13] [14] [15] .Additionally, data mining systems provide the means to easily perform data summarization and visualization, aiding the securityanalyst in identifying areas of concern [16].The models must be represented in some form. Common representationsfor data mining techniques include rules, decision trees, linear and non-linear functions (including neural nets),instance-based examples, and probability models [10].

Classification maps a data item into one of several predefined categories. These algorithms normally output “classifiers”, for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient “normal” and “abnormal” audit data for a user or a program, then apply a classification algorithm to learn a classifier that will determine (future) audit data as belonging

to the normal class or the abnormal class.

Ensemble approaches [4, 6] have the advantage that they can be made to adopt the changes in the stream more accurately than single model techniques. Several ensemble approaches have been proposed for classification of evolving data streams.

Ensemble classification technique is advantageeous over single classification method. It is combination of several base models and it is used for continuous learning. Ensemble classifier has better accuracy over single classification technique. Bagging and boosting are two of the most well-known ensemble learning methods due to their theoretical performance guarantees and strong experimental results. Boosting has attracted much attention in the machine learning community as well as in statistics mainly because of its excellent performance and computational attractiveness for large datasets.

III. PROPOSED APPROACH

This proposed model uses boosted decision tree i.e. hoeffding tree classification techniques to increase performance of the intrusion detection system.

In this proposed scheme boosting method improves ensemble performance by using adaptive window and adaptive size hoeffding tree as base learner. Because of this algorithm works faster and increases performance. It uses dynamic sample weight assignment technique. In this algorithm adaptive sliding window is parameter and assumption free in the sense that it automatically detects and adapts to the current rate of change. Its only parameter is a confidence bound _. Window is not maintained explicitly but compressed using a variant of the exponential histogram technique. It keeps the window of length W using only O (log W) memory & O (log W) processing time per item, rather than the O (W) one expectsfrom a naïve implementation. It is used as change detector since it shrinks window if and only if there has been significant change in recent examples, and estimator for the current average of the sequence it is reading since, with high probability, older parts of the window with a significantly different average are automatically dropped.

Here in our approach we can show the results i.e detection of the intrusions present in one system can be done in another system if the systems are connected via client-server environment via web service (Apache Tomcat) by using the ip address of the client system without installing the software in client system

The underlying idea of boosting is to combine simple rules to form an ensemble such that the performance of the single ensemble member is improved, i.e.boosted..The approach is The underlying idea of boosting is to combine simple rules to form an ensemble such that the

performance of the single ensemble member is improved, i.e. boosted. Let h_1, h_2, …. h_N be a set of hypotheses and consider the composite ensemble hypothesis,

$$f(x)= \sum_{n=1}^{N} \alpha_n h_n \ (x)$$

Here $\_n$ denotes the coefficient with which the ensemble member h_n is combined; both $\_n$ and the learner or hypothesis h_n are to be learned within the boosting procedure.

The boosting algorithm initiates by giving all data training tuples the same weight w_0. After a classifier is built, the weight of each tuple is changed according to the classification given by that classifier. Then, a second classifier is built using the reweighted training tuple. The final classification of a intrusion detection is a weighted average of the individual classifications over all classifiers. There are several methods to update the weights and combine the individual classifiers. After the kth decision tree is built, the total misclassification error _k of the tree, defined as the sum of the weights of misclassified tuples over the sum of the weights of all tuples, is calculated:

$$\varepsilon k=\sum_{i(miscl)} Wi^{k} / \sum_{i} Wi^{k}$$

where i loops over all instances in the data sample. Then, the weights of misclassified tuples are increased

$$W^{k+1} = (1\text{-}\varepsilon k/\varepsilon k)Wi^{k}$$

Finally, the new weights are renormalized as,

$$W^{k+1\text{---->}}\ W^{k+1} / \sum wi^{k+1}$$

and the tree k+1 is constructed. Note that, as the algorithm progresses, the predominance of hard-to-classify instances in the training set is increased. The final classification of tuple i is a weighted sum of the classifications over the individual trees[7, 8, 9]. Furthermore, trees with lower misclassification errors "k are given more weight when the final classification is computed..

In decision tree i.e. hoeffding tree, each node contains a test on an attribute, each branch from a node corresponds to a possible outcome of the test and each leaf contains a class prediction. A decision tree is learned by recursively replacing leaves by test nodes, starting at the root. The attribute to test at a node is chosen by comparing all the available attributes and choosing the best one.

For classifying examples in the dataset,[1] the prior and conditional probabilities generated from the dataset are used to make the prediction. This is done by combining the effects of the different attributes values from the example. Suppose the example e_j has independent attribute values { a_{i1}, a_{i2},…, a_{ip}}, we know $P(a_{ik} \mid c_j)$, for each class c_j and attribute a_{jk} and then estimate $P(e_j|c_j)$ by

$$p(e_i|c_j)=p(c_j)\pi_{k=1\text{--->}p}\ p(a_{ij}|c_j)$$

To classify an example in the dataset, the algorithm estimates the likelihood that e_i is in each class. The probability that e_i is in a class is the product of the conditional probabilities for each attribute value with prior probability for that class. The posterior probability $P(c_j \mid e_i)$ is then found for each class and the example classifies with the highest posterior probability for that example. The algorithm will continue this process until all the examples of sub-datasets or sub-subdatasets are correctly classified[3]. When the algorithm correctly classifies all the examples of all sub or sub-sub datasets, then the algorithm terminates and the prior and conditional probabilities for each sub or sub-sub-datasets are preserved for future classification of unseen examples.

IV. TYPES OF ATTACKS

Now a days many types of attacks are present they are listed in the table below

Main Attacks	**22 Different Attack Types**
DOS	back, land,neptune,pod,smurf,teardrop
U2R	Buffer_overflow,loadmodule,pearl,rootkit
R2L	ftp_write,guess_password,imap,multihop,phf, spy
Probe	Ipsweep,nmap.portsweep,satan

DOS-Denial Of Service:

A denial-of-service attack (DoS attack) or distributed denial-of-service attack (DDoS attack) is an attempt to make a computer or network resource unavailable to its intended users. Although the means to carry out, motives for, and targets of a DoS attack may vary, it generally consists of the concerted efforts of a person, or multiple people to prevent an Internet site or service from functioning efficiently or at all, temporarily or indefinitely.

U2R-User To Root:

A user to root attack is an attempt to hack the data that is being transferred from the user to root.

R2L-Remote To User:

The Remote To User attack can be explained with the help of an example let us see the example below i.e: **ncftp** An R2L attack which exploits a bug in a particular version of ncftp, the popular ftp client. The a user on the victim host ftp's to the attackers machine, and attempts to download, recursively, a directory.

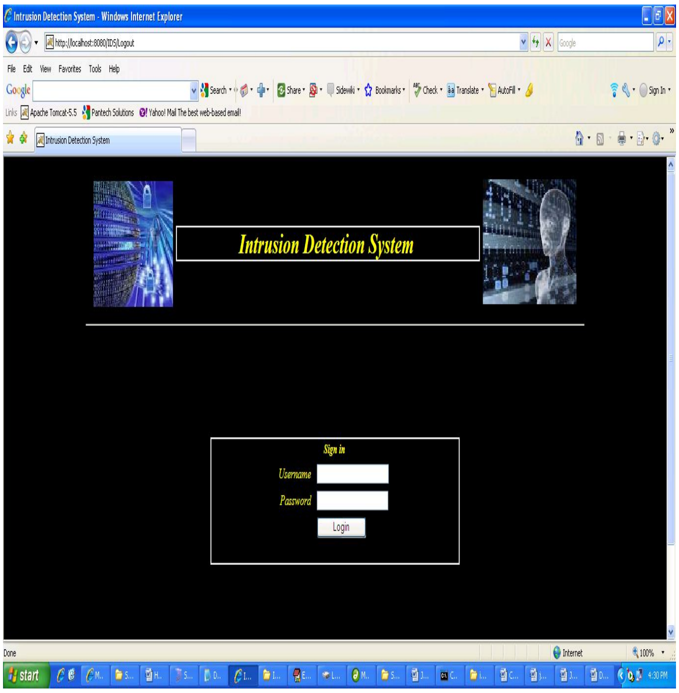
Probe:

A probe is a program or other device inserted at a key juncture in a network for the purpose of monitoring or collecting data about network activity. Relative to computer security in a network, a probe is an attempt to gain access to a computer and its files through a known or probable weak point in the computer system

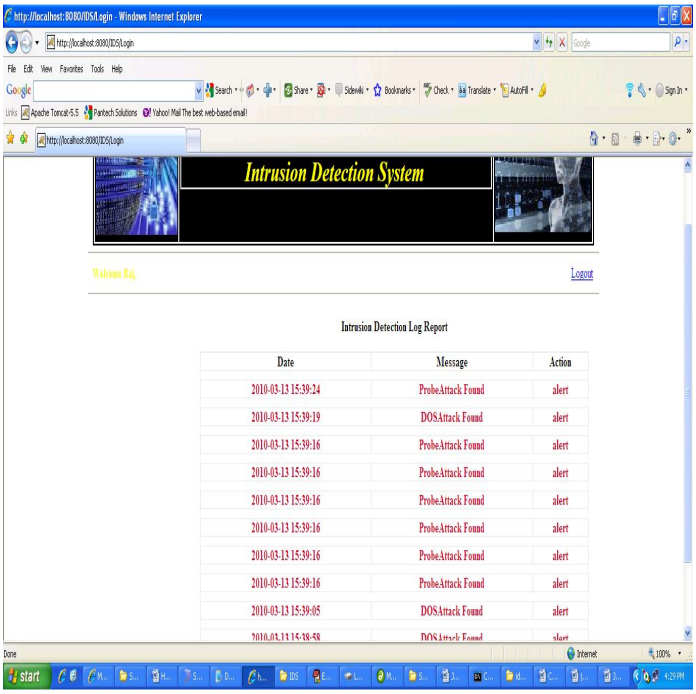
V. EXPERIMENTAL RESULTS

The proposed boosted decision trees algorithm is tested on KDDCup'99 dataset [11] and compared to that of a Naïve Bayes, kNN, eClass0 [2], eClass1 [2] and the Winner (KDDCup'99).

In order to demonstrate the effectiveness of the approach and of the proposed system, we have developed system that asks username and password before showing the results



Seeking username and password


Intrusion Detection System

Welcome Raj,

Logout

Intrusion Detection Log Report

Date	Message	Action
2010-03-13 15:39:24	ProbeAttack Found	alert
2010-03-13 15:39:19	DOSAttack Found	alert
2010-03-13 15:39:16	ProbeAttack Found	alert
2010-03-13 15:39:16	ProbeAttack Found	alert
2010-03-13 15:39:16	ProbeAttack Found	alert
2010-03-13 15:39:16	ProbeAttack Found	alert
2010-03-13 15:39:16	ProbeAttack Found	alert
2010-03-13 15:39:16	ProbeAttack Found	alert
2010-03-13 15:39:05	DOSAttack Found	alert
2010-03-13 15:38:58	DOSAttack Found	alert

Showing the result

The below figures show the graphical representation of different type of attacks like

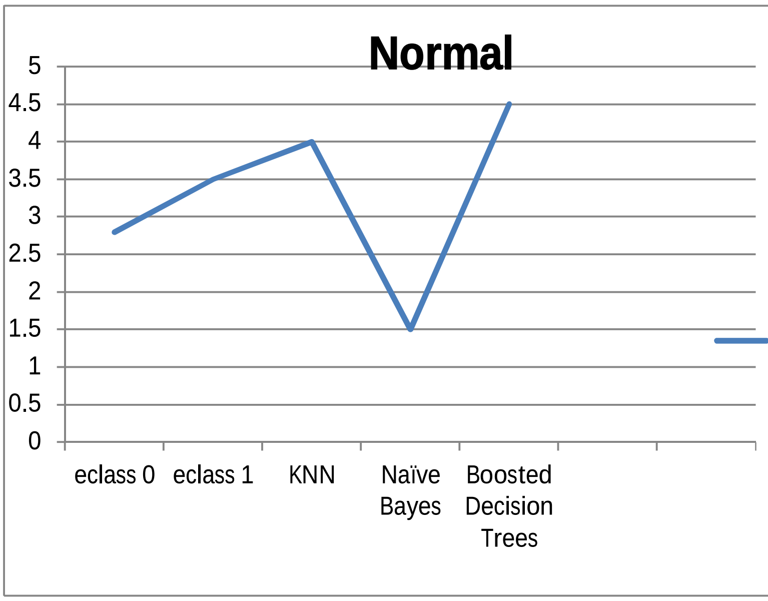
DOS (Denial Of Service)

U2R (User To Root)

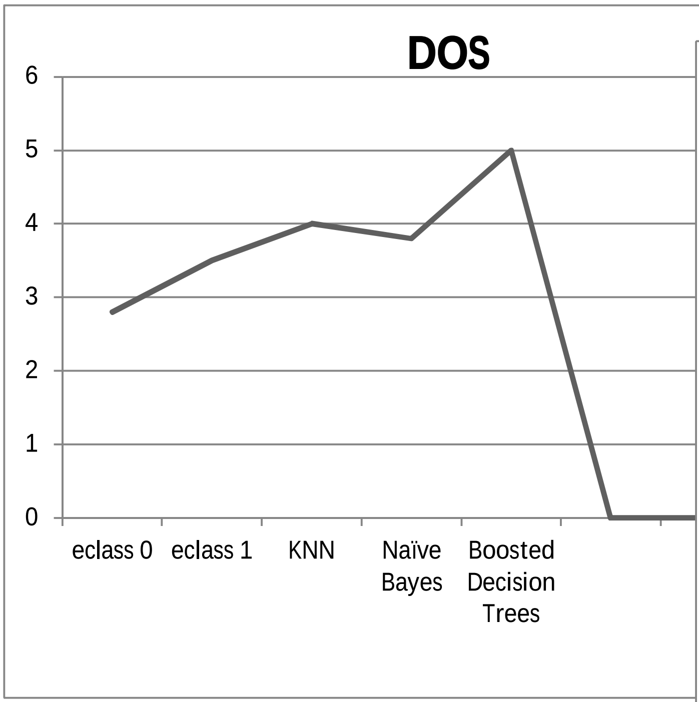
R2L (Remote to User)

Probe

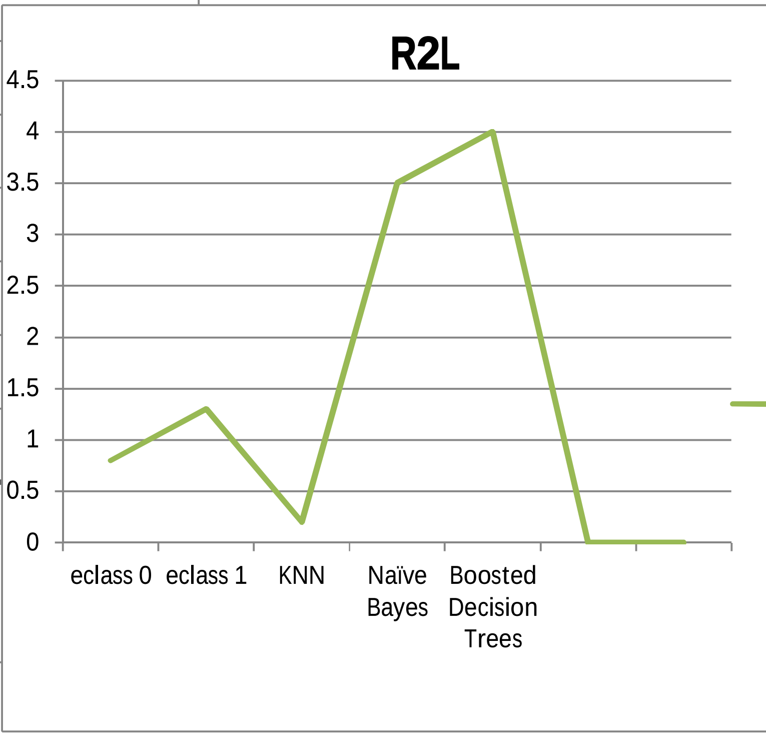
Figures 1(a) - 1(e) show graphical comparison of boosted decision trees algorithm with SVM, kNN and Naïve Bayes with feature selection. 12 features are selected from 41 features.



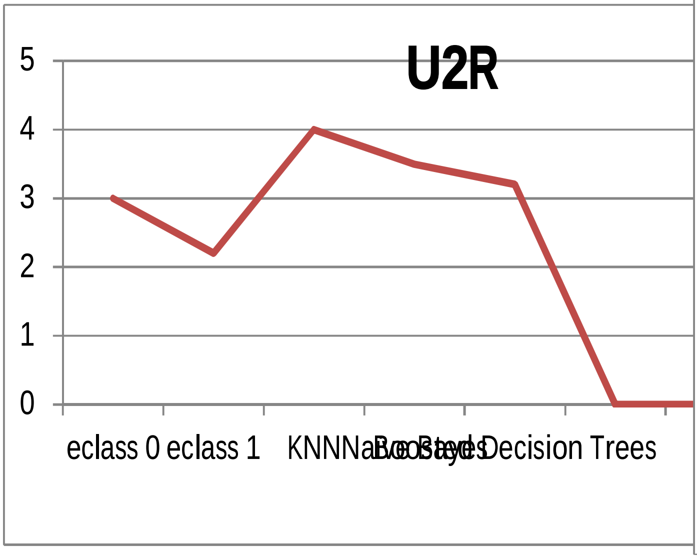
1(a) Normal With 41 Features



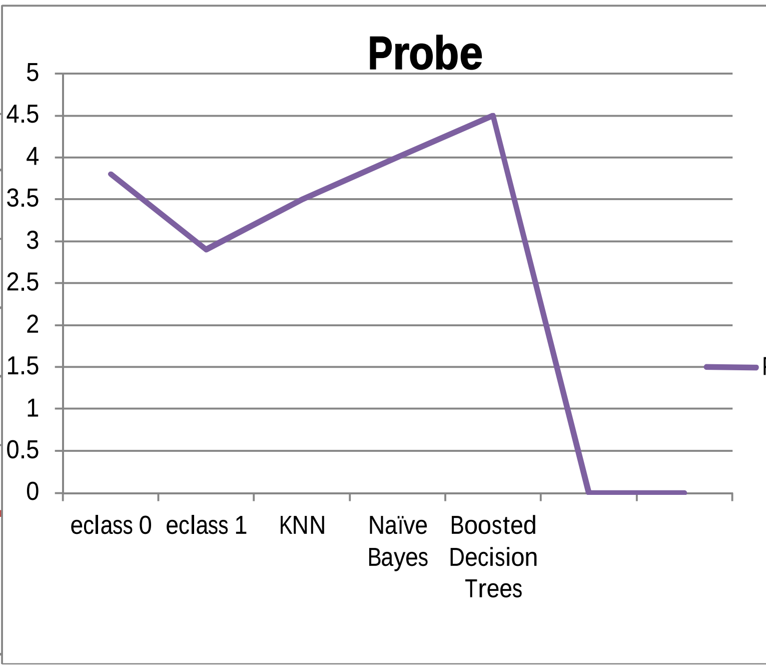
1(b) Dos Attack With 41 Features



1(d) R2L Attack With 41 Features



1(c) U2R Attack With 41 Features



1(e)Probe Attack With 41 Features

VI. CONCLUSION

This paper introduced a network intrusion detection model using boosted decision trees: a learning technique that allows combining several decision trees to form a classifier which is obtained from a weighted majority vote of the classifications given by individual trees.Here boosting is done by combining c45and bayesian algorithm. The generalization accuracy of boosted decision trees has compared with Naïve Bayes, kNN, eClass0, eClass1 and the Winner (KDDCup'99).

On the basis of these results, it can be concluded that boosted decision trees may be a competitive alternative to these techniques in intrusion detection system.

REFERENCES

[1] Daniel Barbara, Ningning Wu and Sushil *Jajodia Detecting novel network intrusion using bayes estimators*. In Proceedings of First SIAM Conference on data mining Chicago, 2001.

[2] Eric Bloedorn, Alan D. Christiansen, William Hill, Clement Skorupka, Lisa M. Talbot, and Jonathan Tivel. *Data mining for network intrusion detection: How to get started*.

[3] E Knorr, Ng, R.: *Algorithms for Mining Distance-based Outliers in Large Data Sets*. Proceedings of the VLDB Conference (1998)

[4] H. Wang, W. Fan, P. Yu, J. Han, "*Mining concept-drifting data streamsusing ensemble classifiers*", InProceedings of the ACM SIGKDD, pp. 226-235, Washington DC, 2003.

[5] M. Masud, J. Gao, L. Khan, J. Han, "*Classifying evolving data streamsfor intrusion detection".*

[6] M. Panda, M. Patra, *"Ensemble rule based classifiers for detecting network intrusions", p*p 19-22, 2009

[7] R. Bane, N. Shivsharan, "*Network intrusion detection system (NIDS)"pp. 1272-1277, 2008*.

[8] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. P. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, M. A. Zissman, "Evaluating *Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation*. *Proceedings DARPA Information Survivability Conference and Exposition*
(DISCEX) 2000", Vol 2, pp. 12--26, IEEE Computer Society Press, Los Alamitos, CA, 2000

[9] S. Ramaswami, R. Rastogi, K.Shim, *"Efficient Algorithms for MiningOutliers from Large Data Sets"*, Proceedings of the ACM SIGMOD Conference, 2000.

[10] S. T. Brugger, "*Data mining methods for network intrusion detection",pp. 1-65, 2004*.

[11] S. Ramaswami, R. Rastogi, K.Shim, *"Efficient Algorithms for MiningOutliers from Large Data Sets*", Proceedings of the ACM SIGMOD Conference, 2000.

[12] V. Barnett and T. Lewis*. Outliers in Statistical Data. John Wiley and Sons, New York*, 1994.

[13] W. Lee, S. J. Stolfo, *"Data Mining Approaches for Intrusion Detection"*, Proceedings of the 1998 USENIX Security Symposium, 1998.

[14] W. Lee, S. J. Stolfo, K. W. Mok, *"A data mining framework for buildingintrusion detection models"*, Proc. of the 1999 IEEE Symp.on Securityand Privacy, pp. 120--132. Oakland, CA, 1999.

[15] Yoav Freund and Robert E. Schapire. *Experiments with a new boosting algorithm. In ICML, pages 148–156, 1996*.

[16] Z. Yu, J. Chen, T. Q. Zhu, *"A novel adaptive intrusion detection systembased on data mining", pp.2390-2395, 2005*