International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal | www.ijtsrd.com
ISSN No: 2456 - 6470 | Volume - 2 | Issue – 6 | Sep – Oct 2018

A Novel Approach of Intrusion Detection System Through SADE in Data Mining

Syed Zuber Hussain[1], Prof. Avinash Sharma[2]
[1]PG Scholar, [2]Assistant Professor
Department of CSE, MITS, Bhopal, Madhya Pradesh, India

ABSTRACT
Intrusion Detection System (IDS) is a vital component of any network in today's world of Internet. IDS are an effective way to detect different kinds of attacks in interconnected network. An effective Intrusion Detection System requires high accuracy and detection rate as well as low false alarm rate. To tackle this growing trend in computer attacks and respond threats, industry professionals and academics are joining forces in order to build Intrusion Detection Systems (IDS) that combine high accuracy with low complexity and time efficiency. With the tremendous growth of usage of internet and development in web applications running on various platforms are becoming the major targets of attack. Security and privacy of a system is compromised, when an intrusion happens. Intrusion Detection System (IDS) plays vital role in network security as it detects various types of attacks in network. Implementation of an IDS is distinguishes between the traffic coming from clients and the traffic originated from the attackers or intruders, in an attempt to simultaneously mitigate the problems of throughput, latency and security of the network.

Data mining based IDS can effectively identify intrusions. The proposed scheme is one of the recent enhancements of naive bayes algorithm. It solves the problem of independence by averaging all models generated by traditional one dependence estimator and is well suited for incremental learning. Empirical results show that proposed model based on SADE is efficient with low FAR and high DR.

Keyword: *Intrusion detection, Data Mining, SADE, NSL-KDD data set, False Alarm Ratio, Detection Rate.*

1. INTRODUCTION
As the years have passed by computer attacks have become less glamorous. Just having a computer or local network connected to the internet, heightens the risk of having perpetrators try to break in, installation of malicious tools and programs, and possibly systems that target machines on the internet in an attempt to remotely control them. The (GOA) team categorized the attacks encountered in 2014 discovering that 25% of the attacks where non-cyber threats followed by scan/probes/attempted access 19% and policy violation 17% [1]. This data is further acknowledged by the annual FBI/CSI survey which discovered that though virus based attacks occurred more frequently, attacks based on un-authorized access and denial of service attacks both internally as well as externally, increased drastically.

Recent exploits also suggest that the more sensitive the information that is held is, the higher the probability of being a target. Several Retailers, banks, public utilities and organizations have lost millions of customer data to attackers, losing money and damaging their brand image [2]. In some cases attackers steal sensitive information and attempt to blackmail companies by threatening to sell it to third parties [5]. In the second quarter of 2014, Code Spaces was forced out of business after attackers deleted its client databases and backups. JP Morgan, Americas‟ largest bank, suffered a cyber-attack in 2014 that impacted 76 million members [3]. In 2014, Benesse, A Japanese Education Company for children suffered a major breach whereby a disgruntled former employee of a third-party partner disclosed up to 28 million customer accounts to advertisers [4]. Most notably the "Sony Pictures hack" best displayed how significant a company's losses are in the aftermath of a security breach. The network servers were

temporarily shut down due to the hack [4]. Cyber Security experts estimate that Sony lost up to $100 million [5] [6]. Other companies under the Sony blanket fell victim to attacks [7]. To tackle this growing trend in computer attacks and respond threat, industry professionals and academics are joining forces in a bid to develop systems that monitor network traffic activity raising alerts for unpermitted activities. These systems are best described as Intrusion Detection Systems.

2. DATA MINING BASED IDS

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Network traffic is huge and information comes from different sources, so the dataset for IDS becomes large. Hence the analysis of data is very hard in case of large dataset. Data mining techniques are applied on IDS because it can extract the hidden in formation and deals with large dataset. Presently Data mining techniques plays a vital role in IDS. By using Data mining techniques, IDS helps to detect abnormal and normal patterns. The various data mining techniques that are used in the context of intrusion detection.

1. **Correlation Analysis:** Correlation is often used as a preliminary technique to discover relationships between variables. More precisely, the correlation is a measure of the linear relationship between two variables.
2. **Feature Selection:** A subset of features available from the data is selected for the application of a learning algorithm.
3. **Machine Learning:** Machine learning explores the study and construction of algorithms that can learn from and make predictions on data
4. **Sequential Patterns:** It is used to excavate connection between data, time series analysis gains more focus on the relationship of data in times.
5. **Classification:** It is a technique of taking each instance of a dataset and assigning it to a particular class. Typical classification techniques are: inductive rule generation, genetic algorithms, fuzzy logic, neural networks and immunological based techniques.
6. **Clustering:** It is a technique for statistical data analysis. It is the classification of similar objects into a series of meaningful subset according to certain rules, so that the data in each subset share some common trait.
7. **Deviation Analysis:** Deviation analysis can reveal surprising facts hidden inside data
8. **Forecast:** Finding certain laws according to historical data, establishing models and predicting types, characteristics of the future data, etc based on the model.

With the increase in computerization and storage of more and more sensitive data on the data servers, the security of the data servers is a major issue. As the intrusion detection systems are being used for monitoring networked devices where they look for the behavior patterns of various anomalous and malicious behaviors in the audit data. Making comprehensive IDS requires more time and expertise. On the other hand Data mining based IDS require less expert knowledge and give better performance (Barbara et al.,2001; Noel et al.,2002; Eskin et al ., 2002 ; Markou and Singh, 2003).They can generalize new and unknown attacks in a better way. The methods used for finding knowledge can be mathematical or non - mathematical; it can be deductive or inductive. The available knowledge can be used for optimizing enquiry, manage information, control progress and make intellectual decision. Given databases of sufficient size and quality, data mining technology can generate.

1. Automated prediction of trends and behaviors:

Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data. It also provides various models that help in forecasting.

2. Automated discovery of previously unknown patterns:

Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions.

3. PROPOSED METHODOLOGY

Optimize Singular Average Dependence Estimators (SADE) is one of the recent enhancements of naive bayes algorithm. OSADE solves the problem of independence by averaging all models generated by traditional one dependence estimator and is well suited for incremental learning. OSADE produces favorable results compared to traditional models. OSADE classifier is widely applied to several problems like bio medical, intrusion detection, spam filtering. OSADE algorithm is proposed to resolve the issues that were identified in ODE and SPODE. Advantages of OSADE algorithm are listed as follows.

- Probabilistic classification learning technique.
- Preferable for data sets where there is dependency among attributes.
- Low variance.
- Predicts class probabilities.
- Accurate and multi class classifier.
- Useful for large data set.

Our proposed approach is shown in figure 2. Our network intrusion detection model applies on the Optimize SADE classifier.

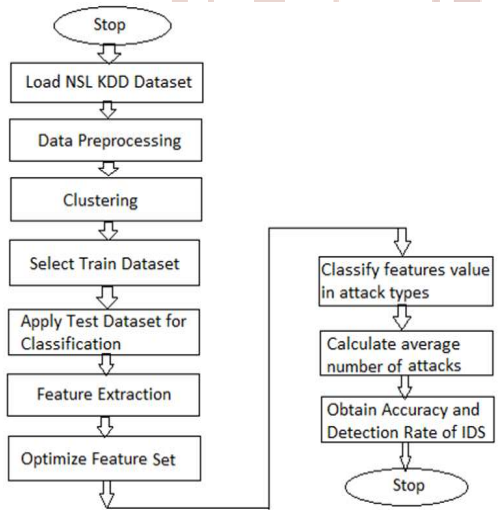

Figure 1: Optimize SADE approach

Our proposed algorithm is described below

Algorithm: Intrusion Detection System using Optimize SADE techniques.

Input: NSL-KDD Data set

Output: Classification of different types of attacks.

Step 1: Load NSL KDD data set.

Step 2: Apply preprocessing technique - discretization.

Step 3: Clustered the datasets into four types.

Step 4: Partition each cluster into training and test sets.

Step 5: Data set is given to proposed algorithm for training.

Step 6: Test dataset is then fed to propose for classification of attacks.

Step 7: Extract the features value of test dataset.

Step 8: Optimize the features value in continuous orthogonal way.

Step 9: Now classify features value as per number of attacks.

Step 10: Determine average number of attacks in respective attack class, namely and DoS, Probe, U2R, R2L.

Step 11: Record the accuracy, detection rate (DR), false alarm rate (FAR), Matthews correlation coefficient (MCC).

4. RESULT ANALYSIS

Many standard data mining process such as data cleaning and pre-processing, clustering, classification, regression, visualization and feature selection are already implemented in MATLAB. The automated data mining tool MATLAB is used to perform the classification experiments on the 20% NSL-KDD dataset. The data set consists of various classes of attacks namely DoS, R2L, U2R and Probe.

The data set to be classified is initially pre-processed and normalized to a range 0 -1. This is done as a requirement because certain classifiers produce a better accuracy rate on normalized data set. Correlation based Feature Selection method is used in this work to reduce the dimensionality of the features available in the data set from 41 to 6. Classification is done in this work by using SADE algorithms.

The specific types of attacks are classified into four major categories. The table 1 shows this detail.

Table 1: Mapping of Attack Class with Attack Type

Attack Class	Attack Type
DoS	Back, Land, Neptune, Pod, Smurf,Teardrop,Apache2, Udpstorm, Processtable, Worm (10)
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint (6)
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named (16)
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps (7)

The Table 2 shows the distribution of the normal and attack records available in the various NSL-KDD datasets.

Table 2: Details of Normal and Attack Data in Different Types of NSL-KDD Data Set

Data Set Type	Parameters					
	Records	Normal Class	DoS Class	Probe Class	U2R Class	R2L Class
KDD Train+ 20%	25192	13449	9234	2289	11	209
		53.39 %	36.65 %	9.09 %	0.04 %	0.83 %
KDD Train+	125973	67343	45927	11656	52	995
		53.46 %	36.46 %	9.25 %	0.04 %	0.79 %
KDD Test+	22544	9711	7458	2421	200	2754
		43.08 %	33.08 %	10.74 %	0.89 %	12.22 %

Figure 2 clearly exhibits the count of normal and various attack class records in the different train and test NSL-KDD data sets.

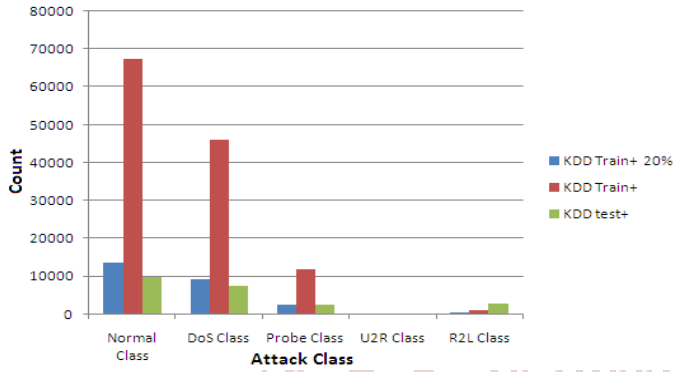

Figure 2: Network vector distribution in various NSL-KDD train and test data set

Further analysis of the KDD Train+ data set has exposed one of the very important facts about the attack class network vectors. From the Figure 2, it is apparent that most of the attacks launched by the attackers use the TCP protocol suite. The transparency and ease of use of the TCP protocol is exploited by attackers to launch network based attacks on the victim computers.

Table 3: Protocols Used By Various Attacks

Attack Class / Protocol	DoS	Probe	R2L	U2R
TCP	42188	5857	995	49
UDP	892	1664	0	3
ICMP	2847	4135	0	0

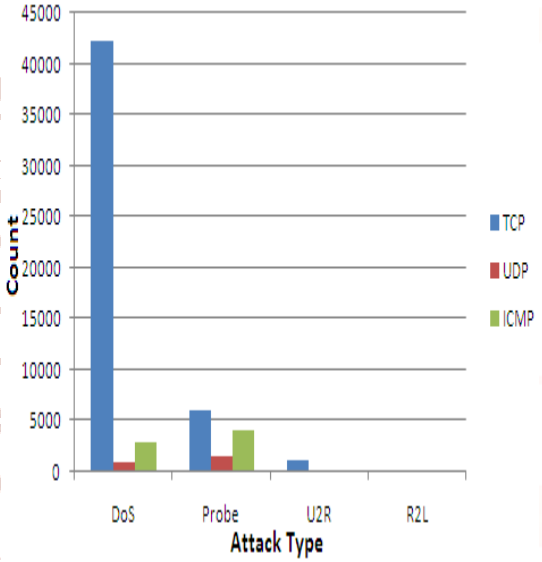

Figure 3: Protocol-wise attacks in the KDD Train+ data set

We used accuracy, detection rate DR), false alarm rate (FAR) and Matthews correlation coefficient (MCC) which are derived using confusion matrix.

Table 4: Confusion Matrix

	Classified as Normal	Classified as Attack
Normal	TP	FP
Attack	FN	TN

Where,
TN -Instances correctly predicted as non-attacks.
FN - Instances wrongly predicted as non-attacks.
FP -Instances wrongly predicted as attacks.
TP -Instances correctly predicted as attacks.
Accuracy = (Number of samples correctly classified in test data)/(Total number of samples in test data)
Detection Rate (DR) = TP / (TP+FN)
False Alarm Rate (FAR) = FP / (FP+TN)
MCC = (TP x TN – FP x FN) / sqrt ((TP+FP)(TP+FN)(TN+FP)(TN+FN)

We conducted all our experiments using WEKA tool [14]. The performance of our proposed model is shown in table 5 and for IIDPS shown in table 6.

Table 5: Performance of our Model

SN	Attack Type	Accuracy	Detection Rate (DR)	False Alarm Rate (FAR)	Matthews Correlation Coefficient (MCC)
1	DoS	97.19	98.63	4.44	0.943
2	Probe	96.48	98.19	5.45	0.927
3	U2R and R2L	96.25	98.65	6.48	0.925

Table 6: Performance of IIDPS[6]

SN	Attack Type	Accuracy	Detection Rate (DR)	False Alarm Rate (FAR)	Matthews Correlation Coefficient (MCC)
1	DoS	89.90	94.85	15.72	0.72
2	Probe	90.48	96.07	15.87	0.812
3	U2R and R2L	90.47	95.60	15.37	0.811

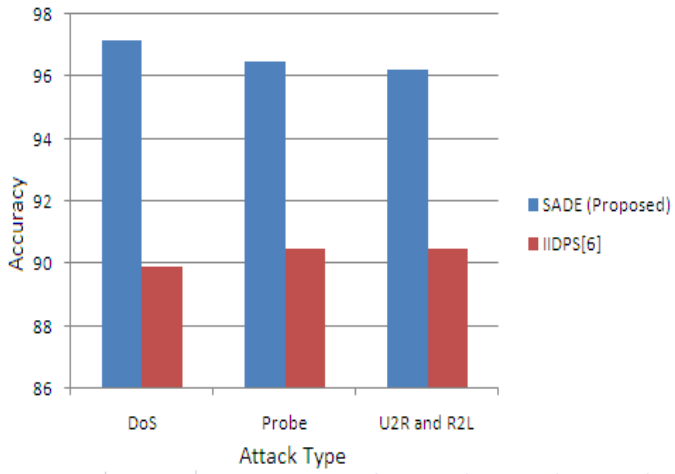

Figure 4: Accuracy of SADE (Proposed) and IIDPS [6]

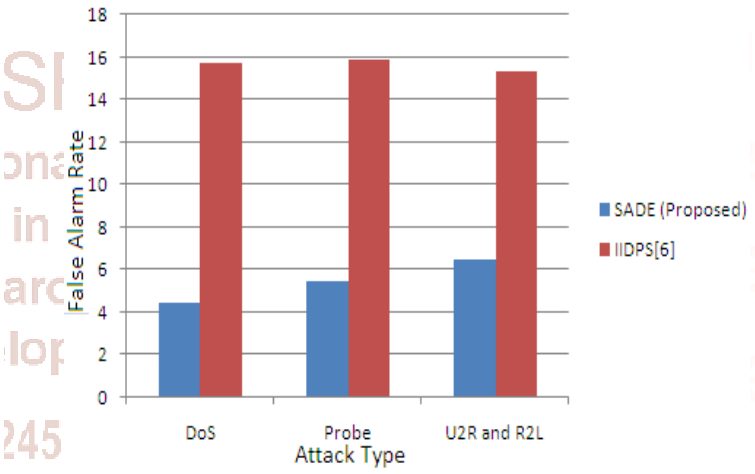

Figure 6: False Alarm Rate of SADE (Proposed) and IIDPS [6]

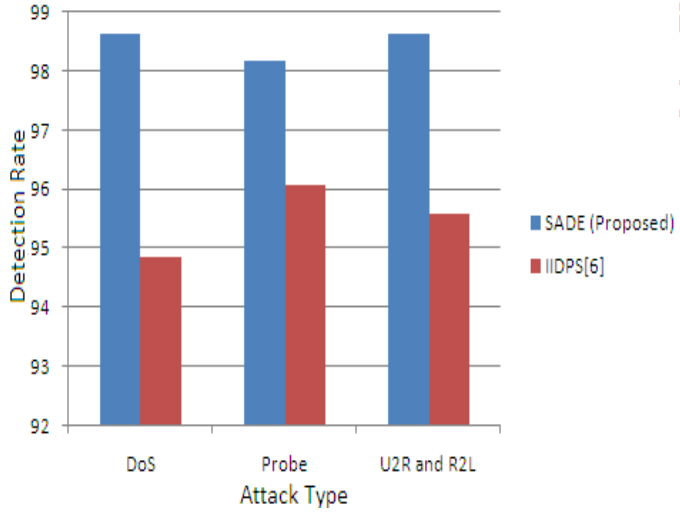

Figure 5: Decision Rate of SADE (Proposed) and IIDPS [6]

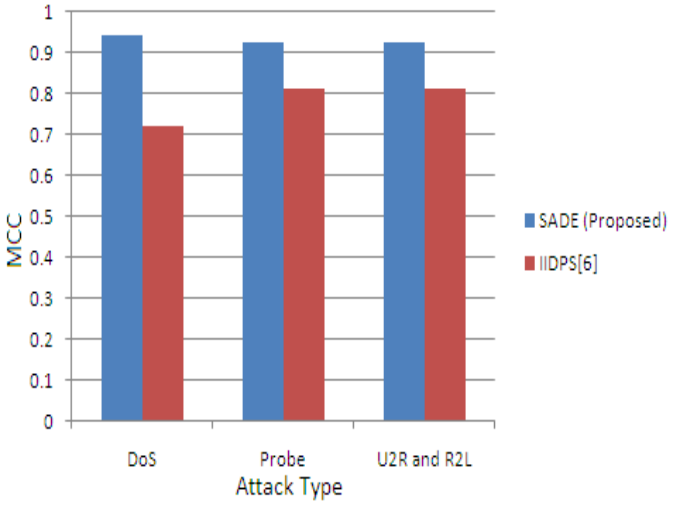

Figure 7: Matthews Correlation Coefficient of SADE (Proposed) and IIDPS [6]

It is evident from tables 5 and 6, figures 4 and 5 that our proposed model yielded high DR and low FAR to classify the attacks. For DOS attack, our proposed model achieved an accuracy of 97.19%, which is 7% more than IIDPS algorithm. FAR recorded for IIDPS is 15.72 which is almost 11% more than our proposed model. For a good classifier to detect attacks it should have high DR and low FAR. For a probe attack FAR is recorded as 15.87% for IIDPS algorithm which is almost 10% more than our proposed model. For R2L and U2R, FAR has been recorded as 15.37% which is almost 8% more than our proposed model. Table 5.11 and figure 1.4 shows accuracy of IIDPS and our proposed model for attack detection.

Mathews correlation coefficient recorded by our model is high compared with IIDPS classifier. Average accuracy recorded by our proposed approach is 96.64%, where as for IIDPS it is only 90.28%. Average value of MCC obtained by our approach is 0.93 and for IIDPS 0.80 only. The experimental result shows that our proposed approach can achieve good accuracy, high DR with low FAR.

5. CONCLUSIONS

In this paper an ANN based Intrusion Detection System was implemented on NSL-KDD dataset. Dataset was trained and tested for binary category (normal or attack) as well as for five class attack categories. Training set having less number of patterns for R2L and U2R categories so some patterns were selected randomly from other three classes in training set. The proposed IDS system uses Levenberg-Marquardt (LM) and BFGS quasi-Newton Backpropagation algorithm for learning. Training and testing applied on dataset with full features (i.e. 41) and with reduced feature (i.e. 29). The result was evaluated based on standard parameter such as accuracy, detection rate and false positive rate and the result was compared with other reported papers. It was found that proposed technique for binary class classification gives higher accuracy of attack detection than that of other reported technique. For five class classification it was found that the system has good capability to find the attack for particular class in NSL-KDD dataset.

In this paper, we applied the SADE algorithm to detect four types of attack like DOS, probe, U2R and R2L. 10 cross validation is applied for classification. The proposed approach is compared and evaluated using NSL KDD data set. Experimental result prove that accuracy, DR and MCC for four types of attacks are increased by our proposed method. Empirical results show that proposed model compared with IIDPS generates low false alarm rate and high detection rate. For future work, we will apply feature selection measure to further improve accuracy of the classifier.

6. REFERENCES

1. Ranju Marwaha, “Intrusion Detection System Using Data Mining Techniques– A Review”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 5, May 2017.
2. Rashmi Ravindra Chaudhari and Sonal Pramod Patil, “Intrusion Detection System: Classification, Techniques and Datasets to Implement”, International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 02, Feb - 2017.
3. Amreen Sultana, M. A. Jabbar, “Intelligent Network Intrusion Detection System using Data Mining Techniques”, IEEE Conference on Data Mining, 2016.
4. Zibusiso Dewa and Leandros A. Maglaras, “Data Mining and Intrusion Detection Systems”, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, 2016.
5. Dikshant Gupta, Suhani Singhal, Shamita Malik and Archana Singh, “Network Intrusion Detection System using various data mining techniques”, International Conference on Research Advances in Integrated Navigation Systems, 2016.
6. Prof. Ujwala Ravale, Prof. Nilesh Marathe and Prof. Puja Padiya, “Feature Selection Based Hybrid Anomaly Intrusion Detection System Using K Means and RBF Kernel Function”, Elsevier Journal, 2015.
7. Solane Duque, Dr. Mohd. Nizam bin Omar, “Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)”, Elsevier Journal, 2015.
8. JABEZ J, Dr. B. MUTHUKUMAR, “Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach”, Elsevier Journal, 2015.
9. D. Shona, A. Shobana, “A Survey on Intrusion Detection using Data Mining Technique”, International Journal of Innovative Research in

Computer and Communication Engineering, Vol. 3, Issue 12, December 2015.

10. Fang-Yie Leu, Kun-Lin Tsai, Yi-Ting Hsiao and Chao-Tung Yang, "An Internal Intrusion Detection and Protection System by Using Data Mining and Forensic Techniques", IEEE SYSTEMS JOURNAL, 2015.
11. G. V. Nadiammai, M. Hemalatha, "Effective approach toward Intrusion Detection System using data mining techniques", Egyptian Informatics Journal, 2014.
12. Kapil Wankhade, Sadia Patka and Ravindra Thool,"An Efficient Approach for Intrusion Detection Using Data Mining Methods", IEEE Conference on Knowledge Engineering, 2013.
13. Kapil Wankhade, Sadia Patka and Ravinrda Thool, "An Overview of Intrusion Detection Based on Data Mining Techniques", International Conference on Communication Systems and Network Technologies, 2013.
14. Reda M. Elbasiony, Elsayed A. Sallam, Tarek E. Eltobely, Mahmoud M. Fahmy "A hybrid network intrusion detection framework based on random forests and weighted k-means", Elsevier Journal, 2013.
15. Muamer N. Mohammad, Norrozila Sulaiman, Osama Abdulkarim Muhsin, "A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment", Elsevier Journal, 2011.
16. D. Powell and R. Stroud, "Conceptual model and architecture", Deliverable D2, project MAFTlA IST-19993-11583,lBM Zurich research laboratory research report R23377, NOV(2011).

IJTSRD
International Journal of Trend in Scientific Research and Development
ISSN: 2456-6470