ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 7, January 2015

Intrusion Detection System using Bayesian Approach

S. Saravanan, Dr. R M. Chandrasekaran
Department of Computer Science & Engineering, Annamalai University
Annamalainagar – 608 002, Tamil Nadu, India.

Abstract*— Today, there is a wide spread of Internet services all over the world, many kinds and large number of security threats are increasing. Since it is not technically feasible to build a system with no vulnerabilities, Intrusion Detection System (IDS), which can effectively detect intrusion accesses have attracted attention. Intrusion detection can be defined as the process of identifying unusual behavior that targets a network and its resources. An ID examines all data features to detect unauthorized or unapproved activity related to Network Security. It plays a vital role for analyzing the network traffic log and each log is characterized by large set of features and it requires huge computational processing power and time for the classification process. So feature selection is important before modeling. This paper aims to identify important reduced features with select by feature Quantile filter and Chi-Squared. The scope of this work is to detect various classes of attacks using Naïve Bayes, Radial Basis and J-48 classifiers are trained and tested individually and the classification rates for different classes are observed. The Naïve Bayes classifier has outperformed well with respect to Accuracy and Classification error rate compared with J-48 and RBF classifier. Empirical results show that selected attributes give better performance to design effective IDS.

***Index Terms*—Intrusion detection systems (IDSs), Quantile filter, Chi-Squared, Principal Component Analysis (PCA), Support Vector Machines (SVM).**

I. INTRODUCTION

Intrusion detection as defined by the Sys Admin, Audit, Networking, and Security (SANS) Institute is the art of detecting inappropriate, inaccurate, or anomalous activity. Today, intrusion detection is one of the high priority and Challenging tasks for network administrators and security professionals. Any intrusion detection system has some inherent requirements. Its prime purpose is to detect as many attacks as possible with minimum number of false alarms; the system must be accurate in detecting attacks. Intrusion detection systems are classified into network based, host based, or application based depending on their mode of deployment and data used for analysis. Intrusion detection systems (IDSs) are software or hardware systems that automate the process of monitoring the events occurring in a computer system or network and analyzing them for signs of security problems. All of them fight against intruders and viruses. When IDS is properly deployed, it can provide warnings indicating that a system is under attack, even if the system is not vulnerable to the specific attack. These warnings can help users alter their installation's defensive position to increase resistance to attack..

II. RELATED WORK

A. Network Intrusion Detection System (NIDS) based on Data Mining, 2013

With the tremendous growth in information technology, network security is one of the challenging issue and so as Intrusion Detection system (IDS). IDS are an essential component of the network to be secured. The traditional IDS are unable to manage various newly arising attacks. To deal with these new problems of networks, data mining based IDS are opening new research avenues. Data mining is used to identify new patterns which were not known previously from large volume of network dataset. New Intrusion Detection Systems are based on sophisticated algorithms in spite of signature based detection [2]. Data mining method uses binary classifiers and multi boosting simultaneously. Features are selected using binary classifiers for more accuracy in each type of attack. Multi boosting is used to reduce both the variance and bias. With data mining, it is easy to identify valid, useful and understandable pattern in large volume of data. Thus the efficiency and accuracy of Intrusion Detection system are increased and security of network so is also enhanced.

B. Using Naïve Bayes Classifier to Accelerate Constructing Fuzzy Intrusion Detection Systems, 2013

A Bayesian classifier [3] is one of the most widely used classifiers which possess several properties that make it surprisingly useful and accurate. It is illustrated that performance of Bayesian learning in some cases is comparable with neural networks and decision trees. Bayesian theorem [1] suggests a straight forward process which is not based on search methods. This is the major point which satisfies the marvelous time complexity of Bayesian classifier. At the other hand, constructing phase of fuzzy intrusion detection systems suffer from time consuming processes which are based on search methods. Authors propose a novel method to accelerate such processes using Bayesian inference.

C. Intrusion Detection System Based on Improved BP Neural Network and Decision Tree, 2012

According to the attributes of both BP Neural Network [4] and Decision Tree, the paper presents an advanced complex-algorithm model in order to improve the ability of intrusion detection. The simulating results show that the new system not only can increase the average detection rate and reduce the failing, but it can also be more effective to

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 7, January 2015

simplify the complexity, raise the detection speed and promote the accuracy by use of abstracting rules and paralleled dealing method in matching process. The experimental results indicate that the intrusion detection rate of the model based on integrated algorithms has no obviously change to some attacks like Neptune, Teardrop and Portsweep, actually reduce slightly, but the detection rate to Guess_passwd has been improved till 92.5%, to Buffer_overflow is up to 45.5%, the whole detection rate has been improve considerably. Based on this, if we add some decision rules appropriately, the detection rate of the model to Buffer_overflow will increase further more. This is the reason why we design in the model to add or modify rules after attacks take place so as to improve its self-learning capability.

D. A Novel Multi-Classifier Layered Approach to Improve Minority Attack Detection in IDS, 2012

Due to the tremendous growth of network based services, intrusion detection has emerged as an important technique for network security. While variety of security techniques are being developed and a lot of research is going on intrusion detection, but the field lacks an integrated approach with high detection rate (recall) and precision for minority attacks namely R2L and U2R. The paper presents a novel layered approach with multi-classifier by combining naïve bayes classifier (NBC) and naive bayes tree (NB Tree) to improve detection rate and precision of minority class without hurting the performance of majority class. To identify important reduced feature set for each attack separately, to form layered approach [5]. They found that a single classifier is not effective in detecting minority attacks with acceptable reliability.

E. A Hybrid Intelligent Approach for Network Intrusion Detection, 2012

Intrusion detection is an emerging area of research in the computer security and networks with the growing usage of internet in everyday life. Most intrusion detection systems (IDSs) mostly use a single classifier algorithm to classify the network traffic data as normal behavior or anomalous. However, these single classifier systems fail to provide the best possible attack detection rate with low false alarm rate. Authors propose to use a hybrid intelligent approach [6] using combination of classifiers in order to make the decision intelligently, so that the overall performance of the resultant model is enhanced. The general procedure in this is to follow the supervised or un-supervised data filtering with classifier or clustered first on the whole training dataset and then the output is applied to another classifier to classify the data. They use 2-class classification strategy along with 10-fold cross validation method to produce the final classification results in terms of normal or intrusion. Experimental results on NSL-KDD dataset, an improved version of KDD Cup 1999 dataset show that our proposed approach is efficient with high detection rate and low false alarm rate.

F. A Comparison of Data Mining Techniques for Intrusion Detection, 2012

The Expositional increase in the traffic across networks has necessitated the need to detect unauthorized access. In this sense Intrusion Detection has become one of the major research areas. The paper three data mining techniques [7] namely CS.O Decision Tree, Ripper Rule and Support Vector Machines are studied and compared for the efficiency in detecting the Intrusion, It is found that the CS.O Decision Tree is efficient than the other two. The data mining tool Clementine is used for evaluating this on the KDD99 dataset [9]. In this work they used Clementine, KDD Dataset via Java programming to collect network packet information including IP header, TCP header, UDP header, and ICMP header from each of the packet, later separates the packet information by considering connections between any two IP address. A number of records have been collected as training and testing datasets for network Intrusion Detection with CS.O Decision Trees, Ripper Rules and SVM.

G. An Approach towards Intrusion Detection using PCA Feature Subsets and SVM, 2012

Presently many intrusion detection approaches are available but have drawbacks like training overhead as well as their performance factor. Increased detection rate with less false alarms can enhanced the efficiency of an intrusion detection system. One of the main limitations is the processing of raw features for classification which increases the architecture complexity and decreases the accuracy of detecting intrusions. Because of the limitations in earlier approaches, this PCA based subsets has been proposed. An SVM based IDS mechanism with Principal Component Analysis (PCA) feature subsets [8] has been presented. Support Vector Machines (SVM) used as classifier to test and train the subsets of extracted features with Radial Basis Function (RBF) kernel. The main issue in intrusion detection system is to design a system having accurate detection of intrusions with fewer false alarms. Apart from this, classifier overhead is another problem because of processing many irrelevant and redundant features which at end causes decrease in detection rate. For this purpose, first authors used PCA in order to transforms the features into high dimension space and then apply traditional reduction technique by extracting different subsets from features vector calculated by PCA. These subsets were used individually to train and test the system and their corresponding results were calculated in terms of sensitivity and specificity. The main focus of this paper is to focus on having possible feature reduction with maximum accuracy using SVM classifier. So keeping these constraints, the subset S10 was selected as best reduced feature subset using PCA traditional technique. There is no complexity as such in the classifier architecture and these reduced features improve classifier performance without having training overhead in processing all features.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 7, January 2015

H. Using Naive Bayes with AdaBoost to Enhance Network Anomaly Intrusion Detection, 2010

Classical intrusion detection system tends to identify attacks by using a set of rules known as signatures defined before the attack; this kind of detection is known as misuse intrusion detection. But reality is not always quantifiable, and this drives us to a new intrusion detection technique known as anomaly intrusion detection, due to the difficulties of defining normal pattern for random data frames, anomaly detection suffer from false positives, where normal traffic behavior is mistaken and classified as an attack and cause a great deal of manpower to manual sort the attacks. Authors construct a network based anomaly intrusion detection system using naive Bayes [13] as weak learners enhanced with AdaBoost [11] (Adaptive Boosting machine learning algorithm), experiment using KDD '99 cup data proved that our IDS can achieve extremely low False Positive and has acceptable detection rate.

I. Host Based Intrusion Detection Using RBF Neural Networks, 2009

A novel approach of host based intrusion detection is suggested in the paper that uses Radial Basis Functions (RBF) Neural Networks [10] as profile containers. The system works by using system calls made by privileged UNIX processes and trains the neural network on its basis. An algorithm is proposed that prioritize the speed and efficiency of the training phase and also limits the false alarm rate. In the detection phase the algorithm provides implementation of window size to detect intrusions that are temporally located. Also a threshold is implemented that is altered on basis of the process behavior. The system is tested with attacks that target different intrusion scenarios. The result shows that the radial Basis Functions Neural Networks provide better detection rate and very low training time as compared to other soft computing methods. The robustness of the training phase is evident by low false alarm rate and high detection capability depicted by the application.

J. Ensembling Rule Based Classifiers for Detecting Network Intrusions, 2009

An intrusion is defined as a violation of the security policy of the system, and hence, intrusion detection mainly refers to the mechanisms that are developed to detect violations of system security policy. Recently, data mining techniques have gained importance in providing the valuable information which in turn can help to enhance the decision on identifying the intrusions (attacks). Authors evaluate the performance of various rule based classifiers [11] like: JRip, RIDOR, NNge and Decision Table using ensemble approach in order to build an efficient network intrusion detection system. To use KDDCup'99, intrusion detection benchmark dataset (which is a part of DARPA evaluation program) for our experimentation. It can be observed from the results that the proposed approach is accurate in detecting network intrusions, provides low false positive rate, simple, reliable and faster in building an efficient network intrusion system..

III. FEATURE SELECTION METHODS

Feature selection can be defined as a process that chooses a minimum subset of M features from the original set of N features, so that the feature space is optimally reduced according to a certain evaluation criterion. As the dimensionality of a domain expands, the number of feature N increases. Finding the best feature subset is usually intractable and many problems related to feature selection have been shown to be NP hard Feature selection is an active field in computer science. It has been a fertile field of research and development since 1970s in statistical pattern recognition machine learning and data mining. Feature selection is a fundamental problem in many different areas, especially in forecasting, document classification, bioinformatics, and object recognition or in modeling of complex technological processes. Datasets with thousands of features are not uncommon in such applications. All features may be important for some problems, but for some target concepts, only a small subset of features is usually relevant. Feature selection [12] reduces the dimensionality of feature space, removes redundant, irrelevant, or noisy data. It brings the immediate effects for application: speeding up a data mining algorithm, improving the data quality and thereof the performance of data mining, and increasing the comprehensibility of the mining results. Feature selection algorithms may be divided into select by feature Quantile filter and Chi-Squared.

i) Select by feature Quantile filter

Quantile filtering selects the most frequently occurring words until the accumulated frequencies exceed a threshold of 1.0. In addition, we include all words from the partition that contributes the word that exceeds the threshold. This filter removes the features which have below 1.0 threshold value.

ii) Chi-Squared

Feature Selection via chi square $\left(X^2\right)$ test is another, very commonly used method. Chi-squared attribute evaluation evaluates the worth of a feature by computing the value of the chi-squared statistic with respect to the class. The initial hypothesis H_0 is the assumption that the two features are unrelated, and it is tested by chi-squared formula:

$$X^2 = \sum_{i=0}^{r} \sum_{j=0}^{c} \frac{(O_{ij} E_{ij})^2}{E_{ij}} \qquad (1)$$

Where O_{ij} is the observed frequency and E_{ij} is the expected (theoretical) frequency, asserted by the null hypothesis. The greater the value of X^2, the greater the evidence against the hypothesis X^2 is shown in (1).

IV. NOVEL LAYERED APPROACH

The goal of using a layered model is to reduce computation and the overall time required to detect anomalous events. The

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 7, January 2015

time required to detect an intrusive event is significant and can be reduced by eliminating the communication overhead among different layers. This can be achieved by making the layers autonomous and self-sufficient to block an attack without the need of a central decision-maker. original datasets can be classified into four layers that correspond to the four attack groups mentioned in the data set. They are Probe layer, DoS layer, R2L layer, and U2R layer.

i) Probe Layer

The probe attacks are aimed at acquiring information about the target network from a source that is often external to the network. Hence, basic connection level features such as the "duration of connection" and "source bytes" are significant while features like "number of files creations" and "number of files accessed" are not expected to provide information for detecting probes.

ii) Dos Layer

The DoS attacks are meant to force the target to stop the service(s) that is provided by flooding it with illegitimate requests. Hence, for the DoS layer, traffic features such as the "percentage of connections having same destination host and same service" and packet level features such as the "source bytes" and "percentage of packets with errors" is significant. To detect DoS attacks, it may not be important to know whether a user is "logged in or not."

iii) R2L Layer

The R2L attacks are one of the most difficult to detect as they involve the network level and the host level features. We therefore selected both the network level features such as the "duration of connection" and "service requested" and the host level features such as the "number of failed login attempts" among others for detecting R2L attacks.

iv) U2R Layer

The U2R attacks involve the semantic details that are very difficult to capture at an early stage. Such attacks are often content based and target an application. Hence, for U2R attacks, we selected features such as "number of file creations" and "number of shell prompts invoked," while we ignored features such as "protocol" and "source bytes."

V. MACHINE LEARNING ALGORITHMS

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.

i) Naïve Bayes Classifier

The naïve Bayes classifier operates on a strong independence assumption. This means that the probability of one attribute does not affect the probability of the other. Given a series of n attributes, the naïve Bayes classifier makes 2n independent assumptions. Nevertheless, the results of the naïve Bayes classifier are often correct. The work reported in examines the circumstances under which the naïve bayes classifier performs well and why. It states that the error is a result of three factors: training data noise, bias, and variance. Training data noise can only be minimized by choosing good training data. The training data must be divided into various groups by the machine learning algorithm. Bias is the error due to groupings in the training data being very large. Variance is the error due to those groupings being too small.

ii) J48 decision Tree

The J48 algorithm is Rapid-miner's implementation of the C4.5 decision tree learner. The algorithm uses a greedy technique to induce decision trees for classification and uses reduced-error pruning.

iii) RBF Classifier (Function)

Radial Basis Function (RBF) network [14] consists of input, hidden and output layer. Each hidden layer node is radial basis function centered on a vector from input space. Output units are weighted sums of the hidden units. We used the Rapid-miner tool for the implementation of RBF nets, which employs K-means algorithm for parameter estimation of Gaussian radial basis functions. Their implementation can work for both types of classes: discrete and numeric.

VI. EXPERIMENTAL RESULTS

In our experiments, each item is described by 41 features which form a vector. Note that some features are continuous and some are nominal. In our work, each model used nearly 17,000 instances of data of KDD Cup 99 dataset [15] for training and testing. To calculate the Classification error rate, this is an estimate of the true error rate and is expected to be a good estimate, if the number of test data is large. It is defined as follows:

Classification error rate =

$$\frac{(Total\ test\ \text{data} - \text{Total correctly classified data})}{\text{Total test data}}$$

$$\text{Accuracy} = \frac{(TP + \text{TN})}{(\text{TP} + \text{FN} + \text{FP} + \text{TN})} * 100$$

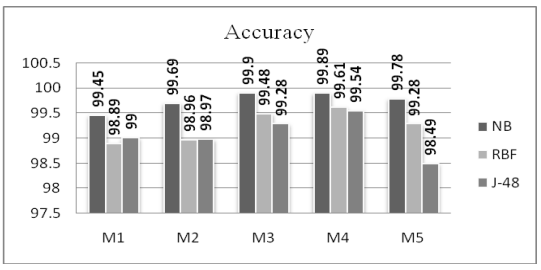

Fig.1 Accuracy of various classifiers using Quantile filter

Fig. 1 shows the accuracy of various classifiers (Naïve Bayes, J-48 and RBF) with Select by feature Quantile Filter. From that figure, the Naïve Bayes classifier shows higher accuracy compare to J-48 and Radial Basis Function (RBF) classifiers.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 7, January 2015

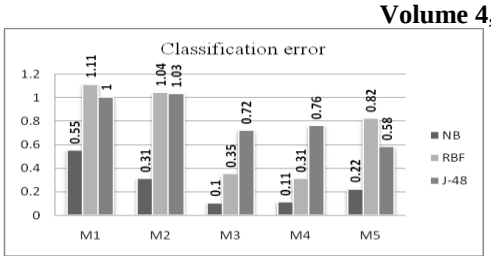

Fig. 2 Classification error of various classifiers using Quantile filter

Fig. 2 shows the classification error of various classifiers (Naïve Bayes, J-48 and RBF) with Select by feature Quantile Filter. From that figure, the Naïve Bayes classifier shows reduced classification error compare to J-48 and Radial Basis Function (RBF) classifiers.

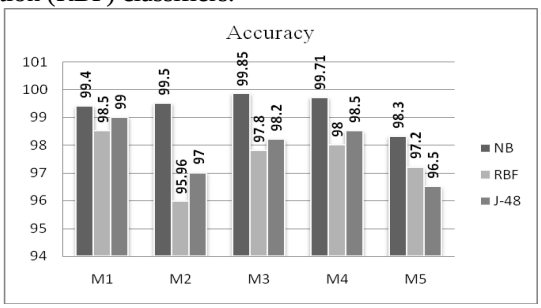

Fig. 3 Accuracy of various classifiers using Chi-Squared

Fig. 3 shows the accuracy of various classifiers (Naïve Bayes, J-48 and RBF) with Chi-Squared. From that figure, the Naïve Bayes classifier shows higher accuracy compare to J-48 and Radial Basis Function (RBF) classifiers.

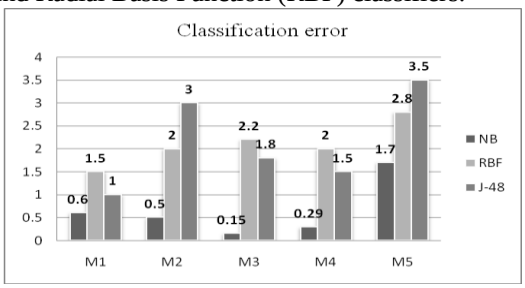

Fig. 4 Classification error of various classifiers using Chi-Squared

Fig. 4 shows the classification error of various classifiers (Naïve Bayes, J-48 and RBF) with Chi-Squared. From that figure, the Naïve Bayes classifier shows reduced classification error compare to J-48 and Radial Basis Function (RBF) classifiers. The above Fig. 1, 2, 3 and 4 depict the performance metrics like Accuracy and Classification error. It is obvious that Naïve Bayes performs well by showing better performance compare to other classifiers. Finally, two feature selection methods are compared, so that we conclude the performance of Naïve Bayes classifier with Select by feature Quantile Filter give better result than Chi-Squared method.

VII. CONCLUSION

Since the goal of this research was to improve the accuracy of the various classes of attacks using Bayesian methods, we have succeeded in achieving our target by using the Bayesian method as an engine to classify the data accordingly. Comparison has been made with three classifiers and the results are shown in above graph. To enhance the results the feature reduction techniques is applied. The feature selection methods are applied to the KDD CUP 1999 dataset to reduce its features and implemented using RapidMiner software. Quantile Filter selects 10 features and Chi-Squared selects 11 features from 41 features data set. The reduced features are used as input to different classifiers and the results are compared. The results show the efficiency with selected features compared to the 41 features, with reduced training and testing times.

VIII. FUTURE WORK

Future work will include customize of feature selection method to improve the results for intrusion particularly for U2R attacks with reduced complexity and overheads.

REFERENCES

[1] Devarakonda N, Pamidi S, Valli Kumari V and Govardhan A " Intrusion Detection System using Bayesian Network and Hidden Markov Model" in Procedia Technology 4 (2012) 506 – 514© 2012 Published by Elsevier Ltd.

[2] S.A.Joshi and Varsha S.Pimprale "Network Intrusion Detection System (NIDS) based on Data Mining" In International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013.

[3] Mehran Amiri, et.al. "Using Naïve Bayes Classifier to Accelerate Constructing Fuzzy Intrusion Detection Systems" in International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[4] Jinhua Huang and Jiqing Liu "Intrusion Detection System Based on Improved BP Neural Network and Decision Tree" in 2012 IEEE fifth International Conference on Advanced Computational Intelligence (ICACI) October 18-20, 2012.

[5] Neelam Sharma and Saurabh Mukherjee" A Novel Multi-Classifier Layered Approach to Improve Minority Attack Detection in IDS" In Procedia Technology 6 (2012) 913 – 921© 2012 Published by Elsevier Ltd.

[6] Mrutyunjaya Panda., et.al, "A Hybrid Intelligent Approach for Network Intrusion Detection" In Procedia Engineering 30 (2012) 1 – 9© 2012 Published by Elsevier Ltd.

[7] R.China Appala and **P.S.Avadhani "A Comparison of Data Mining Techniques for Intrusion Detection" in IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2012.

[8] Noreen Kausar, et.al, "An Approach towards Intrusion Detection using PCA Feature Subsets and SVM" in International Conference on Computer & Information Science (ICCIS), 2012.

[9] Wei Li and QingXia Li "Using Naive Bayes with AdaBoost to Enhance Network Anomaly Intrusion Detection" in Third International Conference on Intelligent Networks and Intelligent Systems, 2010.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 7, January 2015

[10] Usman Ahmed and Asif Masood "Host Based Intrusion Detection Using RBF Neural Networks" in International Conference on Emerging Technologies, 2009.

[11] Mrutyunjaya Panda and Manas Ranjan Patra "Ensembling Rule Based Classifiers For Detecting Network Intrusions" in International Conference on Advances in Recent Technologies in Communication and Computing, 2009.

[12] Jasmina Novaković, et.al, "Toward optimal feature selection using ranking methods and classification algorithms" In Yugoslav Journal of Operations Research 21 (2011), Number 1, 119-135 DOI: 10.2298/YJOR1101119N.

[13] Dewan Md. Farid and Mohammad Zahidur Rahman "Adaptive Intrusion Detection based on Boosting and Naïve Bayesian Classifier "in International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.

[14] Dilip Kumar Ahirwar, Sumit Kumar Saxena and M. S. Sisodia "Anomaly Detection by Naive Bayes & RBF Network" In International Journal of Advanced Research in Computer Science and Electronics Engineering, Volume 1, Issue 1, March 2012, ISSN: 2277 – 9043.

[15] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set" Proceeding of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Application (CIDA 2009)

AUTHOR BIOGRAPHY

Mr. S. Saravanan received the B. E. degree in Computer Science and Engineering from Annamalai University in 1998. He received the M.E. degree in Computer Science and Engineering from Annamalai University, Annamalai Nagar in the year 2005. He has been with Annamalai University, since 2005. He is doing his Ph.D in Computer Science and Engineering at Annamalai University. He published several papers in international conferences and journals. His research interest includes Computer Networks, Intrusion Detection System, Network Security, Mobile Ad hoc Networks and Network Simulator.

Dr. RM. Chandrasekaran is currently working as a Director, Directorate of Distance Education, Annamalai University and Professor at the Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamilnadu, India. From 1999 to 2001 he worked as a software consultant in Etiam, Inc, California, USA. He received his Ph.D degree in 2006 from Annamalai University, Chidambaram. He has conducted workshops and conferences in the area of Multimedia, Business Intelligence, Analysis of Algorithms and Data Mining. He has presented and published more than 45 papers in conferences and journals and is the co-author of the book Numerical Methods with C++ Program (PHI, 2005). His research interests include Data Mining, Algorithms and Mobile Computing. He is life member of the Computer Society of India, Indian Society for Technical Education, Institute of Engineers and Indian Science Congress Association.

.