Ensemble Design For Intrusion Detection Systems

T. Subbulakshmi[1], A. Ramamoorthi[2], and Dr. S. Mercy Shalinie[3]

[1]Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai
subbulakshmitce@yahoo.com
[2]IVCSE, Computer Science Department, Sethu Institute of technology, Madurai,
armoorthi@gmail.com
[3]HODCSE, Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai
shalinie_m@yahoo.com

Abstract

Intrusion Detection problem is one of the most promising research issues of Information Security. The problem provides excellent opportunities in terms of providing host and network security. Intrusion detection is divided into two categories with respect to the type of detection. Misuse detection and Anomaly detection. Intrusion detection is done using rule based, Statistical, and Soft computing techniques. The rule based measures provides better results but the extensibility of the approach is still a question. The statistical measures are lagging in identifying the new types of attacks. Soft Computing Techniques offers good results since learning is done using the training, and during testing the new pattern of attacks was also recognized appreciably. This paper aims at detecting Intruders using both Misuse and Anomaly detection by applying Ensemble of soft Computing Techniques. Neural networks, Support Vector Machines and Naïve Bayes Classifiers are trained and tested individually and the classification rates for different classes are observed. Then threshold values are set for all the classes. Based on this threshold value the ensemble approach produces result for various classes. The standard kddcup'99 dataset is used in this research for Misuse detection. Shonlau dataset of truncated UNIX commands is used for Anomaly detection. The detection rate and false alarm rates are notified. Multilayer Perceptrons, Naïve Bayes classifiers and Support vector machines with three kernel functions are used for detecting intruders. The Precision, Recall and F- Measure for all the techniques are calculated. The cost of the techniques is estimated using the cost measures. The Receiver Operating Characteristic (ROC) curves are drawn for all the techniques. The results show that Support Vector Machines and Ensemble approach provides better detection rate of 99% than the other algorithms.

Keywords

Intrusion Detection Systems, Anomaly Detection Systems, Misuse Detection Systems, Support Vector Machines, Naïve Bayes Classifiers, Multilayer Perceptrons, Ensemble approach

1. Introduction

Intrusion detection can be defined as "the process of identifying and responding to malicious activity targeted at computing and networking resources". The process of intrusion detection involves both detection tools and people. An Intrusion Detection System(IDS) is a software tool used to detect unauthorized access to a computer system or network. Intrusion detection is also defined as the most popular way to detect intrusions is by using the audit data generated by the operating systems. Since almost all activities are logged on a system, it is possible that a manual inspection of these logs would allow intrusions to be detected. It important to analyze the audit data even after an attack has occurred to determine the extent of damage sustained this analysis also helps in tracking down the attacks and in recording the attacks patterns for future detection. A good Intrusion Detection that can be used to analyze audit data for insights makes a valuable tool for information systems. The two types of intrusion detection systems are Anomaly detection system and misuse detection system. In this paper we are going to create a ensemble approach for anomaly and misuse detection system and use the approach as model and compare the result of anomaly detection system and misuse detection system for different parameters using the Soft computing techniques.

2. Review Of Literature

In the research done by Dong Seong Kim et. al[1] 1999 KDD intrusion detection contest data set is used. KDD 1999 dataset was preprocessed and learning & testing is done. In learning process they used various C (1,500,1000) values, and kernal functions linear, polynomial, RBF. The set contains 14 new attacks which is used as test and evaluate the model. They use 4898431 instances as training set and 311029 instances as test set. Tabulate the values for various kernal function and C values for validation experiment and got 93.56% for linear and testing experiment.

In the research done by Nahla Ben Amor et. al[2] 1999 KDD intrusion detection contest data set is used. Naive bayes classifier composed of two levels: one root node which represents a session class(normal and different types of attacks) , several leaf nodes, and several sub leafs each of them contains a feature of connection. Decision tree classification is also discussed in this paper, it has three basic component “selecting decision node” specifying test attribute, “selecting edge” or branch corresponding to the one of the possible attribute values which means one of the test attribute outcomes. “leaf” which is also named an answer node, contains the class to which the object belongs. In naive bayes classification a graphical component is used which composed of a directed acyclic graph where vertices represent events and edges are relations between events, and numerical component consisting in a qualification of different links in the DAG by conditional probability distribution. and focuses on three cases. In each of the studied cases, the evaluation of classification efficiency is based on the percent of correct classification and tabulated the result for cases and confusion matrices.

In the research done by Wilson Naik Bhukya et. al[3] schonalu data set is used. Usually certain cast formulations have been used to compute the overall performance of masquerade detection methods. In this paper a new formula is proposed for effectiveness of masquerading detection and also highly effective approach to masquerader detection using Hidden Markov Models (HMM). Overall goodness of a masquerade detection algorithm was achieved by creating a scoring formulation. According to new formulation cost of masquerade detection define as

$$\text{Cost} = 6 * \text{FPR} + (100 - \text{DR}).$$

The authors felt that this approach reduced the FPR and proposed a new formula for ranking the methods in masquerade detection by calculating its overall effectiveness. The result are tabulated with all the previous approaches with all performance factors. There should be more focus on maximum DR with low FPR for efficient masquerade detection approach.

$$\text{Effectiveness} = (1-\alpha) * \text{DR} + \alpha * \text{FPR}$$

This formula might appear simplistic but it gives a weight to the detection rate. HMMs are initialized and trained. Thresholds of HMM are used for detecting normal behavior. Finally it is concluded that training of HMMs is computationally expensive. The result shows a 90.9% of DR and low FPR of 8.07%. The overall effectiveness of the approach is 74.33.

In the research done by Yingbing Yu et. al[4] schonalu data set is used. This paper uses the finite automata based model to construct a normal behavior reference model from the analysis of shell command sequences. A fuzzy evaluation mechanism is proposed to classify the degree of threat as linguistic terms. The fuzzy number calculated from the output of a fuzzy inference system compared with predefined generalized fuzzy numbers representing different threat levels. This paper presented a finite automata based model to build the normal behavior reference model from shell command sequences generated by UNIX/Linux systems. New activities from a user will be compared with the finite automaton to determine the match and mismatch. Three rules have been chosen for the finite automaton to estimate the two values for different scenarios. Experimental result of detection and false alarm rate was tabulated. This method achieves a detection rate of 87.1% and a missing rate of 12.9%

In the research done by Wun-Hwa et. al[5] DARPA data set is used.In this paper two data mining methodologies, artificial Neural networks (ANNs) and Support Vector Machine (SVM) and two encoding methods, simple frequency-based scheme and tf × idf scheme are used to detect potential system intrusion and results show that SVM with tf × idf scheme achieved the best performance. The data used in experiments are BSM audit data from the DARPA 1998 intrusion detection evaluation program at MIT's Lincoln Labs. In this paper BSM audit data are converted and represented by the frequency distribution of the system calls. The training data set is then separated into attack data sets and normal data sets, which are then subsequently fed into the ANN and SVM algorithms. The values are tabulated for both model and attack detection rate of SVM result was reached to 100% and false-positive rate of 10%. The ANN attack detection rate reached 100% with a false positive rate of about 40.72%. ROC curves for ANN and SVM models with the tf × idf encoding method and also for different encoding methods are showed. Finally the authors concluded that there is a need for new ways to identify attacks.

In the research done by Animwsh Patcha et. al [7] provide a comprehensive survey of anomaly detection systems and hybrid intrusion detection systems of the recent past and present and also discussed. Recent technological trends in anomaly detection are briefed and open problems and challenges in this area were identified.

In the research done by Roy A.Maxion et. al[8] schonalu data set is used.An important issue in Masquerader Detection Systems is obtaining audit data. An immense contribution in this regard has been made by schonlau in extending classification with a new algorithm, a 56% improvement in masquerade detection was achieved at a corresponding false-alarm rate of 1.3%.

The review paper by schonlau[9] et al, compared the performance of six masquerade detection algorithms. Researchers target a false alarm rare of 1%2. all methods had relatively low hit rates(39.4% - 69.3%) and high false alarm rates (1.4% - 6.7%). The results were compared using both cluster and ROC curves, revealing that no single method completely dominated any other. An overview of the various masquerade detectors used by schonlau et al are uniqueness , bayes 1-step markov, hybrid multi-step markov, compression ,IPAM , sequence-match. Error analysis for the base results of the 1v49 experiment are: 62.8% hits, 37.2% misses, and 4.63 false alarms.

In the research done by Min Yang et. al[10] schonalu data set is used. This paper proposes a new method of masquerade detection based on string kernel. String kernel is an inner product in the feature space generated by all subsequences of length k. by using string kernel, OCSVM can directly process the UNIX command sequences, which are the input data of masquerade detection, for PU dataset the detection rate of our method is improved by 15% compared with the other unsupervised methods, given the same false positive rate for SEA dataset, our method can achieve about the same detection rate as the best supervised method; compared with the RBF-OCSVM and detection rate is improved by about 13%.. ROC curves for detection rate vs false positive rate, detection accuracy vs window size and detection rate vs step and also false positive rate vs detection rate are displayed. Finally concluded that OCSVM is a good unsupervised algorithm which best fits the problem. RBF-OCSVM detection rate is improved by about 13%.

3. Problem Definition

The Goal of this research work is to apply the Soft Computing Techniques to the Intrusion Detection Systems using the Standard datasets and to determine the Detection Rate and False Positive Rate. The Machine Learning algorithms are applied to both the categories of Intrusion detection Systems : anomaly detection system and misuse detection systems. The Machine Learning algorithms used in this research are ANN, NBC and SVM. The Standard datasets used are Shonlau's[13] standard truncated command line dataset for Anomaly detection systems and kddcup'99 dataset [14] for Misuse detection systems.

4. Dataset

4.1. Anomaly Detection Dataset

The Audit source for anomaly detection experiments is derived from the Truncated command sequences of Schonlau. Data set is collected with seeded masquerading users to compare various intrusion detection methods. The data set consist of 50 records representing to one user each. Each record contains 15,000 commands. The audit data is generated with unix_acct. The first 5000 commands for each user do not contain any masqueraders and are intended as training data. The next 10,000 commands can be thought of as 100 blocks of 100 commands each. This is represented by 100×50 matrix and seeded with masquerading users, i.e. with data of another user not among the 50 users. There is a windows ascii file, of size 100x50. Each column corresponds to one of the 50 users. Each row corresponds to a set of 100 commands, starting with command 5001 and ending with command 15000. The entries in the files are 0 or 1. 0 means that the corresponding 100 commands are not contaminated by a masquerader. 1 means they are contaminated.

4.2. Misuse Detection Dataset

The 1998 DRAPA Intrusion Detection evaluation program was prepared and managed by MIT Lincoln labs. The objective was to survey and evaluate research in intrusion detection. A standard set of data to be audited, which includes a wide variety if intrusions simulated in a military network environment, was provided the 1999 KDD intrusion detection contest uses a version of this dataset.In the 1998 DARPA intrusion detection evaluation program, an environment Was set up to acquire raw TCP/IP dump data for a network by simulating a typical U.S. Air Force LAN. The LAN was operated like true environment, but being blasted with multiple attacks. For each TCP/IP connection, 41 various quantitative and qualitative features were extracted. Of this database a subset of 494021 data were used, of which 20% represent normal patterns. There are 41 features in each record of network intrusion datasets (KDDCUP). The four different categories of attack patterns are DOS, U2S,R2L and Probe

5. System Design

5.1. Pre-processing

Preprocessing is the process of converting raw data into machine input. SVM requires that each data instance is represented as a vector of real number. Hence, if there are categorical attributes, we first have to convert raw data set into numeric data. In the data pre-processing we converted the entire data set into SVMs format it means convert all attribute of the record into real numbers. In SVM preprocessing schonlau data set is converted in to 50 training files and 50 testing file which is used for anomaly detection system and KDDCUP dataset is converted into various training and testing file each training and testing file contains different numbers of record. ANN and Naive bayes requires that each of training or testing file in the format of name of attribute followed by attributes list. In SVM preprocessing truncated command data set is converted in to one file which contain 50 user commands, first 66 percentage of the data is consider as training data and remaining data is consider as testing data for anomaly detection and for misuse detection KDDCUP dataset is used which contains 41 features, KDDCUP dataset have features like protocol name and some other names these non-numeric values are converted into numeric values and divided in to various number of data files, first 66 percentage of the data is consider as training data and remaining data is consider as testing data for misuse detection systems.

5.2. Detection Model

The Intrusion detection Model using the ensemble approach is depicted in fig.1. The training and testing of various soft computing techniques are shown.

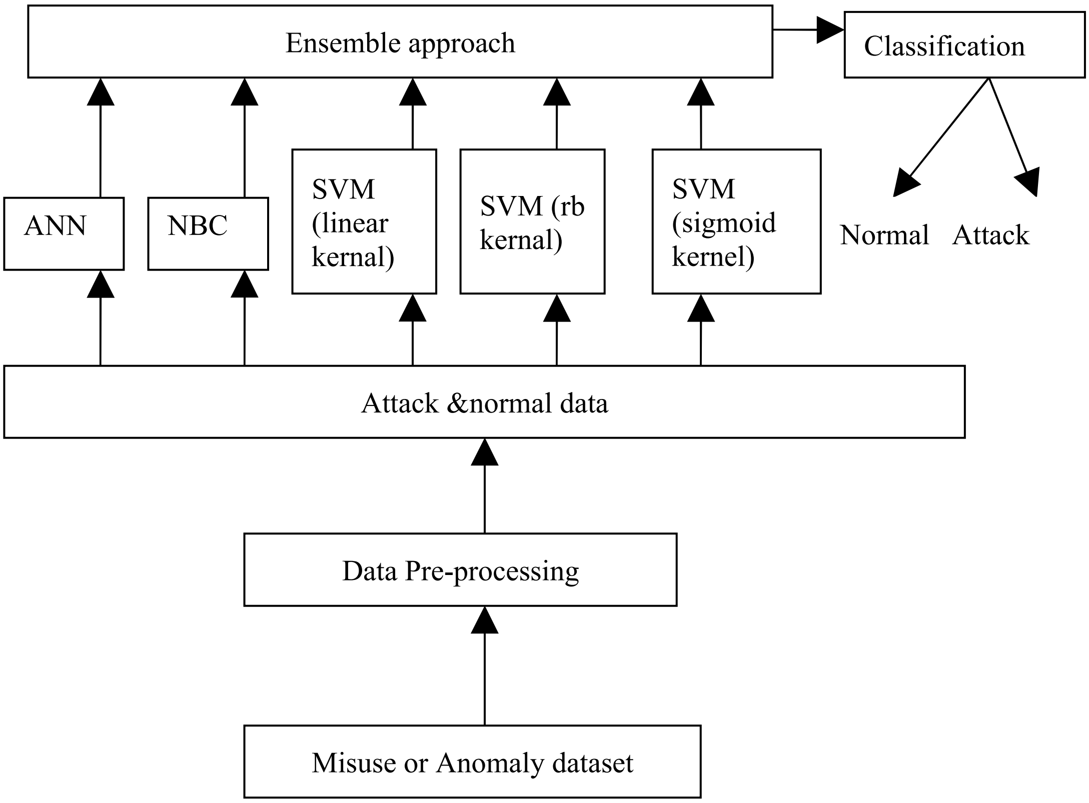

Fig: 1 Ensemble IDS Model

The Misuse or anomaly detection raw data has been pre-processed and converted into training and testing dataset according to type of intrusion detection. The pre processed data is given to the Artificial Neural Networks, Naïve Bayes Classifiers, Support vector Machines(linear kernal), Support vector machines(Radial Basis Function kernal), Support Vector Machines (Sigmoid kernal). Individually these models will produce detection results for various attack and normal classes. Based on the classification results of the Individual Models, the threshold values are assigned for various classes in the range[0,1]. For example if Artificial Neural Networks best classifies DOS attack classes, then the threshold value of DOS attacks for Artificial Neural Networks is set to be high[0.99] and for other classifiers like Naïve Bayes and Support Vector Machines it is set to be very low. This is given as input to the ensemble approach. So an attack will be finally classified as DOS attack only if it is been classified by ANN.

6. Results and Analysis

Table:1 Detection results for Anomaly Detection Total no of instances: 7679(50 user), Total no of training instances:5068(34 user), Total no of testing instances:2611 (16 user)

Name of the classifier	**Correctly classified instances**	**Incorrectly classified instances**	**Classification Accuracy**	**Time for classification**
ANN	2514	97	96.28	12.92
NBC	2514	97	96.28	0.64
SVM(linear)	2532	79	96.97	0.58
SVM(rb)	2529	82	96.86	0.02
SVM(sigmoid)	2526	85	96.74	0.08
Ensemble approach	2602	9	99.66	0.32

The classifiers and their corresponding classification rates are listed in Table.1. From the available 50 users the training and testing instances are selected randomly. Some users will be having pure non masquerading command blocks during training. So if the training is done with pure non masquerading blocks the detection of masquerading command blocks in the testing phase is not proper. For these types of users some of the masquerading command blocks in testing files are swapped with the non masquerading command blocks in the training file and the training and testing is done with this swapped version of files to have good classification results. From the Table1. it is observed that the ensemble approach gives good classification accuracy with appreciable time of classification. The performance metrics for these classifiers are listed in the Table 2.

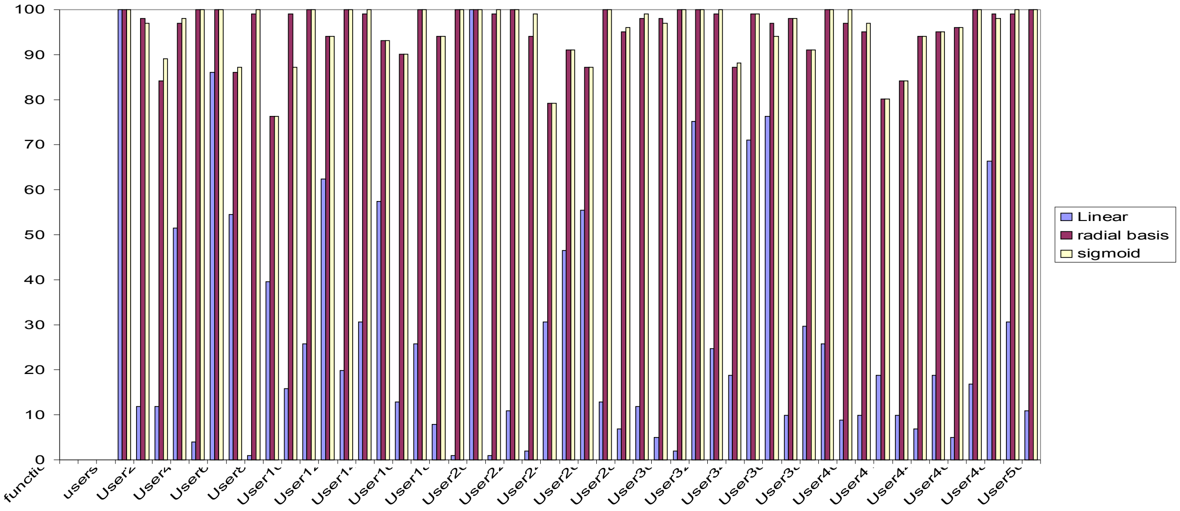

Fig.2. Performance of various kernal functions for the 50 users.

The Table.2 lists the performance metrics Precision Recall and F – Measure. The ensemble approach attains good values in all aspects. The Fig 2. Shows the results of three kernal functions linear, polynomial and Sigmoidal for all the 50 users. The colors of the graph indicates various kernal functions and their performances

Table:2 Performance metrics for Anomaly Detection

Name of the classifier	Precision	Recall	F- Measure
ANN	0.963	0.934	0.917
NBC	0.971	0.942	0.971
SVM(linear)	0.963	0.911	0.981
SVM(rb)	0.911	0.928	0.934
SVM(sigmoid)	0.962	0.976	0.912
Ensemble approach	0.945	0.911	0.952

Table:3 Detection results for Misuse Detection (Total no of instances: 2693, Total no of training instances:1777, Total no of testing instances:916)

Name of the classifier	Correctly classified instances	Incorrectly classified instances	Classification Accuracy	Time for classification (secs)
ANN	887	29	96.83	0.09
NBC	836	80	91.26	2.13
SVM(linear)	891	25	97.27	1.37
SVM(rb)	836	80	91.26	0.09
SVM(sigmoid)	854	62	93.23	0.15
Ensemble approach	900	16	98.25	0.13

Table:4 Performance Metrics for Misuse Detection

Name of the classifier	Precision	Recall	F-Measure
ANN	0.641	0.657	0.627
NBC	0.648	0.621	0.630
SVM(linear)	0.606	0.593	0.589
SVM(rb)	0.640	0.657	0.627
SVM(sigmoid)	0.621	0.669	0.596
Ensemble approach	0.606	0.623	0.789

Table:5 Detection result of SVM for Misuse Detection for various number of input records

Total no of records	No of training records	No of testing records	No of new attacks	Classification Accuracy			
				Linear	polynomial	rb	sigmoid
997	361	643	1	86.93	53.34	72.47	64.54
1043	361	682	7	81.96	50.29	68.32	60.85
1492	993	499	6	74.34	0	72.94	72.94
1977	1310	667	3	92.80	0	92.50	92.40
2030	1350	680	2	83.82	25.14	64.85	63.52
2693	1792	901	2	88.56	18.97	82.79	82.79

7. Conclusion

Thus this paper uses Misuse and Anomaly detection using SVM , NBayes, ANN and ensemble approach. The performance among the various kernal function such as linear, radial basis, and sigmoid were analyzed. Classification accuracy of radial basis and sigmoid kernal functions were founded to be better for anomaly detection and linear kernal function was founded to be better for misuse detection. Ensemble approach outperforms all the other approaches with high classification rate. In the future we will evaluate the performance for the remaining kernal functions in SVMs such as frame, htrbf, and wavelet. And best kernal function among them for misuse and anomaly detection system will be found.

8. References

[1] Dong Seong Kim and Jong Son Park, " Network Based Intrusion Detection with Support Vector Machines", ICOIN 2003, LNCS 2662

[2] Nahla Ben Amor, Salem Benferhat, Zied Elouedi "Naive Bayes vs. Decision trees in intrusion detection systems", Sas'04, March14-17, 2004, Nicosia,Cyprus

[3] Wilson Naik Bhukya, Suresh Kumar G , Atul Negi " A Study of Effectiveness in Masquerade Detection", 2006 IEEE.

[4] Yingbing Yu, James H. Graham, Member, IEEE "Anomaly Instruction Detection of Masqueraders and Threat Evaluation Using Fuzzy Logic", 2006 IEEE

[5] Wun-Hwa Chen, Sheng-Hsun Hsu* , Hwang-Pin Shen "Application of SVM and ANN intrusion detection" , 2004 Elseiver

[6] Kanchan Thadani, Aahutosh, V.K.jayaraman and V.Sundarajan "Evolutionary Selection of Kernels in Support Vector Machines", 2006 IEEE

[7] Animwsh Patcha, Jun "An overview of anomaly detection techniques: Existing solution and latest technological trends" , 2007 Elsevier B.V

[8] Roy A.Maxion, Tahlia N.Townsend " Masquerade Detection Augmented With Error Analysis" ,2004 IEEE

[9] Taeshik Shon , Jongsub Moon " A hybrid machine learning approach to network anomaly detection " 2007 Elsevier Inc.

[10] Min Yang, Huang Zhang , H.J. Cai "Masquerade Detection Using String Kernels " , 2007 IEEE.

[11] Kunlun Li ,Guifa Teng " Unsupervised SVM Based on p-kernels for Anomaly Detection" 2006 IEEE

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

[13] http://www.schonlau.net/

[14] http://www.sigkdd.org/kddcup/index.php?section=1999&method=data

[15] Matthias Schonlau, William Du Mouchel, Wen-Hua. Ju, Alan F. Karr, Martin, Theus and Yehuda vardi, "Computer Intrusion: Detecting Masqueraders", Statistical Science Journal, 2001

[16] Kwong H Yung, "Using Self- Consistent Naïve Bayes classifiers to detect masqueraders", Standard EC Journal, 2004

[17] Mizuki Oka, Yoshihiro Oyama, Hirotake abe, Kazuhiko kato, "Anomaly Detection using layered Networks based on Eigen Co-occurance Matrix", Recent Advances in Intrusion Detection 2004, 2004

[18] Alen peacock, Xian KE, Mattiiew, Wilkerson, "Typing patterns: A key to User Identification", IEEE Security and Privacy, 2004.

[19] Robert Birkely, 2003, “A Neural Network based Intelligent Intrusion Detection System”, M. S. Thesis, Griffith University, Gold coast campus, 2003

[20] Zhuowei Li , Amitabha Das and Jianying Zhou “Theoretical Basis for Intrusion Detection”, 2005 IEEE

Authors

T. Subbulakshmi is working as a Senior Grade Lecturer in department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, Tamilnadu. She has completed her B. E from Raja College of Engineering and Technology, TamilNadu and M.E from Arulmigu kalasalingam College of Engineering, TamilNadu She has published papers in conferences and Journals. She is currently pursuing Ph. D in the area of Information Security. Her research interests includes information Security and Machine learning algorithms

Ramamoorthi is the B.E Computer Science

Student. Pursuing his BE degree in Sethu Institute of Technology, AnnaUniversity, TamilNadu. He has published papers in conferences. His research interests includes Soft Computing Techniques, Network Security and Intrusion Detection Systems

Dr. S. Mercy Shalinie, is currently heading the Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, Tamilnadu. She has has published 50 papers in International Journals. Her research interests includes Application of Neuro Fuzzy systems to various research problems.