IJCSN International Journal of Computer Science and Network, Volume 4, Issue 3, June 2015
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 0.417

A Review on Hybrid Intrusion Detection System Using TAN & SVM

[1] Sumalatha Potteti, [2] Namita Parati

[1]Assistant Professor, Department of CSE, BRECW, Hyderabad, India

[2]Assistant Professor, Department of CSE, BRECW, Hyderabad, India

Abstract - The dramatically development of internet, Security of network traffic is becoming a major issue of computer network system. Attacks on the network are increasing day-by-day. The Hybrid framework would henceforth, will lead to effective, adaptive and intelligent intrusion detection. In this paper, We propose a hybrid fuzzy rough with Naive bayes classifier, Support Vector Machine and K-nearest neighbor (K-NN) based classifier (FRNN) to classify the patterns in the reduced datasets, obtained from the fuzzy rough bioinspired algorithm search. The proposed hybrid is subsequently validated using real-life datasets obtained from the University of California, Irvine machine learning repository. Simulation results demonstrate that the proposed hybrid produces good classification accuracy. Finally, parametric and nonparametric statistical tests of significance are carried out to observe consistency of the classifiers.

Keywords - *Intrusion Detection System (IDS), Data Mining, Classification, Support vector machines (SVM), K-Nearest Neighbor (KNN), Naive Bayes Classifier.*

1. Introduction

IDS is the area, where Data mining is used extensively, this is due to limited scalability, adaptability and validity. In IDS data is collected from various sources like network log data, host data etc. Since the network traffic is large, the analysis of data is too hard. This give rise to the need of using IDS along with different Data mining techniques for intrusion detection. Classifier is more efficient in case of known attacks but for unknown vulnerabilities it gives low detection rate.

1.1 Intrusion Detection System

An IDS is a combination of software and hardware which are used for detecting intrusion[2]. Intrusions may be defined as the unauthorized attempt for gaining access on a secured system or network. Intrusion detection is the course of action to detect suspicious activity on the network or a device. Intrusion Detection System (IDS) is an important detection used as a countermeasure to preserve data integrity and system availability from attacks. The IDS has been a renowned aspect for detecting intrusions adequately. The IDS is assumed as hardware or software or combination of both that allows monitoring of the network traffic in search of intrusions. An intrusion detection system (IDS) inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system It gathers and analyzes the network traffic & detect the malicious patterns and finally alert to the proper authority. The main function of IDS includes:[14]

1.1.1 Classification of IDS

According to techniques used for intrusion detection based on whether attack's patterns are known or unknown, IDS classified into two category [6][15]:
(1) Misuse detection
(2)Anomaly detection

1) Misuse Detection:
Misuse detection compares the user activities to the known intruder activities on web. The idea of misuse detection is to represent attacks in the form of a pattern or a signature so that the same attack can be detected and prevented in future [3]. The IDS searches for defined signatures and if a match is found, the system generates an alarm indicting the presence of intrusion. Since it works on the basis of predefined signatures, it is unable to detect new or previously unknown intrusions.

2)Anomaly Detection:
Anomaly intrusion detection identifies deviations from the normal usage behavior patterns to identify the intrusion [4]. It is a technique which is based on the revealing of traffic anomalies. It estimates the deviation of a user

activity from the normal behavior and if the deviation goes beyond a preset threshold, it considers that activity as an intrusion. It is because of this threshold concept anomaly can detect new intrusions in addition to the previously known intrusions. However anomaly is able to detect new intrusion but the compulsion for involvement of limiting factor results in high percentage of false positive rate.

2. Data Mining Based Intrusion Detection System

Data mining is the activity of extracting relevant information from a large amount of data[17]. Network traffic is massive and information comes from different sources, so the dataset for IDS becomes large. Hence the analysis of data is very shard in case of large dataset. Data mining techniques are applied on IDS because it can extract the hidden information 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

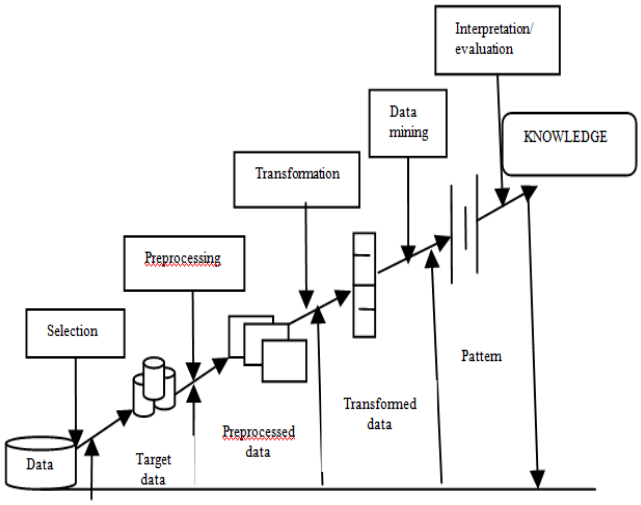

Fig 1: Data Mining

patterns. This section describes different Data mining techniques such as clustering and classification, which are used in IDS to obtain information about vulnerability by monitoring network data[2].

2.1 Classification [2]

Classification is the task of taking each and every instances of dataset under consideration and assigning it to a particular class normal and abnormal means known structure is used for new instances. It can be effective for both misuse detection and anomaly detection, but more frequently used for misuse detection. Classification categorized the datasets into predetermined sets. It is less efficient in intrusion detection as compared to clustering. Different classification techniques such as Naive Bayes classifier, Support Vector Machine and K-nearest neighbor classifier decision tree algorithms are described below:

2.1.1 Naive Bayes Classifier [13]

Naive Bayes classifier is probabilistic classifier. It predicts the class according to membership probability. To derive conditional probability, it analyzes the relation between independent and dependent variable.

2.1.2 Support Vector Machine [12]

Support Vector Machine is supervised learning method used for prediction and classification. It separate data decision tree has high detection rate in case of large points into two classes +1 and -1 using hyperplane because it is binary classification classifier. +1 represents normal data and -1 for suspicious data. Hyperplane can be expressed as: W. X +b=0 Where W={w 1 ,w 2 ,.....,w n } are weight vector for 'n' attributes A={A 1 ,A 2 ,.....,A n }, X={x 1 ,x 2 ,.....,x n } are attribute values and b is a scalar. The main goal of SVM is to find a linear optimal hyper plane so that the margin of separation between the two classes is maximized. The SVM uses a portion of the data to train the system.

2.1.3 K-Nearest Neighbor [11]

It is one of the simplest classification technique. It calculates the distance between different data points on the input vectors and assigns the unlabeled data point to its nearest neighbor class. K is an important parameter. If k=1, then the object is assigned to the class of its nearest neighbor. When value of K is large, then it takes large time for prediction and influence the accuracy by reduces the effect of noise.

2.2 Clustering [2]

Since the network data is too huge, labelling of each and every instances or data points in classification is expensive and time consuming. Clustering is the technique of labelling data and assign into groups of similar objects without using known structure of data points. Members of same cluster are similar and instances of different clusters are different from each other.

Clustering technique can be classified into four groups: Hierarchical algorithm, Partitioning algorithm, Grid based algorithm and Density based algorithm. Some clustering algorithms are explained here.

2.2.1 K-Means Clustering Algorithm [18][13]

K-Means clustering algorithm is simplest and widely used clustering technique proposed by James Macqueen. In this algorithm, number of clusters K is specified by user means classifies instances into predefined number of cluster. The first step of K-Means clustering is to choose k instances as a center of clusters. Next assign each instances of dataset to nearest cluster. For instance assignment, measure the distance between centroid and each instances using Euclidean distance and according to minimum distance assign each and every data points into cluster. K –Means algorithm takes less execution time, when it applied on small dataset. When the data point increases to maximum then it takes maximum execution time. It is fast iterative algorithm but it is sensitive to outlier and noise.

2.2.2 K-Medoids Clustering Algorithm [13]

K-Medoids is clustering by partitioning algorithm as like as K-means algorithm. The most centrally situated instance in a cluster is considered as centroid in place of taking mean value of the objects in K-Means clustering. This centrally located object is called reference point and medoid. It minimizes the distance between centroid and data points means minimize the squared error. K-Medoids algorithm performs better than K-Means algorithm when the number of data points increases to maximum. It is robust in presence of noise and outlier. Generally IDSs are deployed to monitor a system or a network in search of any abnormal condition. In this surveillance if any kind of intrusive attempt is detected, the monitoring system i.e. IDS sets up an alarm which is an indication of the presence of intrusion. In order to detect intrusions in an efficient manner, various appreciable models have registered their presence in the literature.

The presently available models involve usage of various novel algorithms which are likely to detect these intrusions distinguishably. Among these, algorithms based on data mining have been a point of attraction for researchers because of their extensive feasibility in detecting intrusions. These algorithms aid in improving accuracy of the system along with effective detection rate and less false alarm rate. The algorithms loyal for classification are the most desirable algorithms for detection. In the data mining classification techniques, Tree Augmented Naïve Bayes (TAN) and Reduced Error Pruning (REP) algorithms have come out as the most significant detection algorithms in IDS. Hence this paper presents an intelligent effort for intrusion detection which proposes a framework named Hybrid Intrusion Detection Model. This model is a combinational scheme which aims at surmounting the shortcomings faced by two algorithms individually with interestingly increased accuracy of the detection.

3. Proposed Methodology

The proposed system (shown in figure 2) is a hybrid intrusion detection framework based on the combination of two classifiers i.e. Tree Augmented Naïve Bayes (TAN), Support Vector Machine(SVM) . The TAN classifier is used as a base classifier while the SVM classifier is used as a Meta classifier.

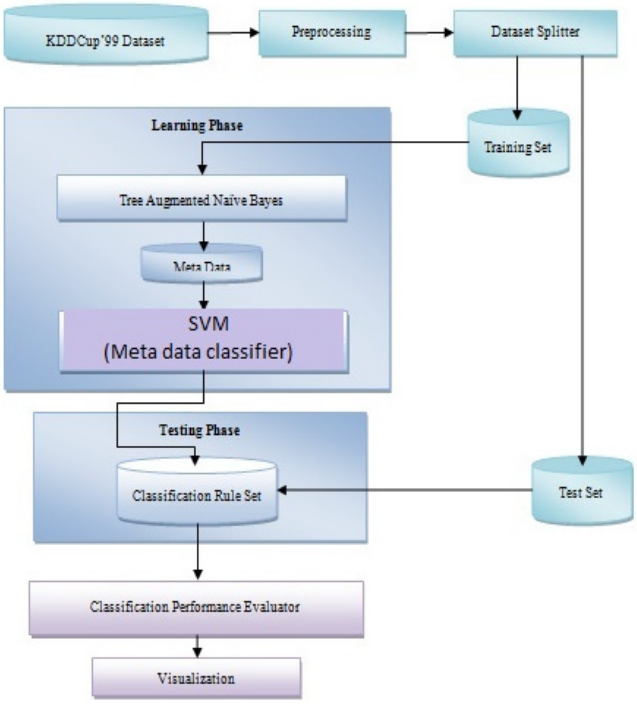

Fig 2: Naive Bayes and SVM

The Meta classification is the learning technique which learns from the Meta data and judge the correctness of the classification of each instance by base classifier. The judgment from each classifier for each class is treated as a feature, and then builds another classifier, i.e. a meta-classifier, to make the final decision [19]. Hence it can be said that the Meta-classification re-classifies the classification judgments made by classifiers. The main idea of using this technique is to improve the overall classification performance resulting in better outcomes than any other existing technique. The two classifiers indulged in the proposed system can be understood as:

3.1Tree Augmented Naïve Bayes Algorithm

The Tree Augmented Naïve Bayes (TAN) [20, 21] is a Bayesian Network learning technique and it is the extension to simple Naïve Bayes classifier. Naive Bayes is probabilistic classifier structure based on Bayes theorem having naive (strong) independence assumptions. This

structure encodes the strong conditional independence assumption among attributes i.e. the class node is the parent node for each and every attribute node with no parent node defined for it.

3.2 SVM Classifier

SVM is developed on the principle of structural risk minimization.

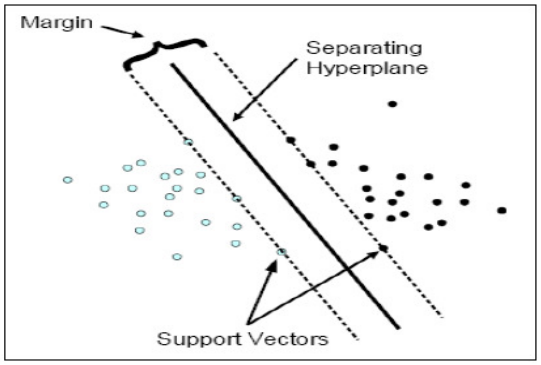

Fig 3: Separating Hyperplabe with SVM

It is one of the learning machines that map the training patterns into the high-dimensional feature space through some nonlinear mapping. SVM has been successively applied to many applications in the multiclass classification [11]. By computing the hyper plain of a given set of training samples, a support vector machine builds up a mechanism to predict which category a new sample falls into figure 3.

4. Detailed Description of the Hybrid IDS Framework

This section describes about all the modules incorporated in the Hybrid IDS framework shown in fig. 2. Following is the brief discussion about each module:

4.1 KddCup'99 Dataset

The kddcup'99 dataset [5] is a benchmark dataset which is originated by processing the tcpdump segment of DARPA 1998 evaluation dataset. The KddCup'99 dataset was originated by processing the tcpdump segment of DARPA 1998 evaluation dataset. The data set consists of 41 features and a separate feature (42nd feature) that labels the connection as 'normal' or a type of attack. The data set contains a total of 24 attack types that fall into 4 major categories (DoS, Probe, R2L and U2R) that are already discussed. For the training and testing of the proposed framework the 10% of the KddCup'99 dataset is used as the full KddCup'99 dataset consists of 5 million instances many of them are redundant. The 10% of the KddCup'99 dataset consists of 494021 instances. In which 97278 are 'Normal' instances and remaining 396743 are belongs to any one type of attack.

4.2. Preprocessing

In the preprocessing module the class label presents in the 42nd feature of KddCup'99 dataset is recast into five major categories for the sake of decreasing complexity of performance evaluation of the proposed model. As the original KddCup'99 dataset having 22 types of attack labels, it was very inconvenient to assess the performance of the classification model. Hence the attack labels are modified to their respective categories for the ease of analysis. Finally five major classes are formed as the class label i.e. DoS, Probe, R2L, U2R and Normal.
The four different categories of attack patterns are as follows.

-**Probing Attack:** It is a method of gathering information about a network of computers with an intention of circumventing its security controls.

-Denial of Service Attack (DoS): It is a type of attack in which an attacker denies legitimate users access to machines or makes computing resources too busy to handle requests.

Table 1. Four types of attacks in KddCup'99 Dataset

Attack type	Attack name
Prob ing	ipsweep nmap por tswee p satan
DoS	back land neptune pod smu rf teardrop
U2R	rootkit perl loadm odu le buffer-over flo w
R2L	ftp-write spy phf guess-passwd imap warezclien t warezmaster multihop

-User to Root (U2R): In U2R the attacker first accesses the system with a normal user account by sniffing passwords or social engineering and then gains root access to the system by exploiting some vulnerability.

-Remote to Local (R2L): R2L occurs when a user without an account has the ability to send packets to a machine gains local access as a user of that machine. Table 1 shows four types of attacks in KddCup'99 Dataset.

4.3. Dataset Splitter

The Dataset Splitter module partitions the dataset into two parts received from the preprocessing module. To partition the dataset into two parts a method named holdout is used. In this method, the given data are randomly partitioned into two independent sets, a training set and a test set [17]. The 66% of the data is allocated to the training set and the remaining 44% of the dataset is allocated to the testing set. The training set is used to derive the proposed framework while the test set is used to assess the accuracy of the derived model. When the KddCup'99 dataset passed through the data splitting module then it gets divided into the training set which consists of 326054 instances and the testing set which consists of 167967 instances.

4.4. Learning Phase

The learning phase involves two steps for generating the classification rules. In the first step, the learning of base classifier i.e. TAN using the training dataset is achieved. The outcome of this base classifier is assumed as the input data (known as Meta data) for the second step. This meta-level training set is composed by using the base classifiers predictions on the validation set as attribute values, and the true class as the target [18]. From these predictions, the meta-learner adapts the characteristics and performance of the base classifier and computes a meta-classifier which is a model of the original training data set. This meta-classifier in second step fetches the predictions from the base classifier for classifying an unlabeled instance, and then makes the final classification decision.

4.5. Testing Phase

The classification rules that are generated in Learning Phase are stored for the performance evaluation of hybrid intrusion detection framework. In this phase, the Testing Set generated in Data Splitting module is used as input to assess the performance. The outcomes of this module is further forwarded to next module i.e. Classifier Performance Evaluator module.

4.6. Classifier Performance Evaluator:

The Classifier Performance Evaluator module calculates the various classification performance

• True Positive Rate (TPR): $$TPR = \frac{TP}{TP + FN}$$

• False Positive Rate (FPR): $$FPR = \frac{FP}{TN + FP}$$

• **True Negative (TN):** These are the negative tuples that were correctly labeled by the classifier.
• **True Positive (TP):** These refer the positive tuples that were correctly labeled by the classifier.
• **False Positive (FP):** These are the negative tuples that were incorrectly labeled as positive.
• **False Negative (FN):** These are the positive tuples that were mislabeled as negative.

4.7. Visualization

The result generated in the Performance Evaluation phase can be visualized in the visualization module**.** These results can be in the form of text or graph etc.

This paper proposes an envisioning framework for intrusion detection i.e. Hybrid Intrusion Detection System. The developed framework is an intelligent, adaptive and effective intrusion detection framework. The experimental analysis is performed on the developed IDS framework and is compared with other techniques present in the scenario. The resultants obtained convey that the developed hybrid framework is highly effective to overcome the deficiencies found in previous work. As the framework uses two data mining techniques (i.e. TAN& SVM) to breed the classification rules, it can be effortlessly implemented in real time and is able to detect and adapt new types of intrusive activities. Also experimental assessment shows that the developed framework has reduced the false alarm rate and increased the accuracy up to noteworthy extend which is a major concern in case of intrusion detection mechanism. In addition to this, the framework is able to detect U2R and R2L attacks more efficiently than previous findings, boosting up the detection process. Based on whether a learning algorithm is included in the training process or not, existing feature selection (FS) approaches can be broadly classified into two categories: filter and wrapper approaches. In wrapper approaches, a learning algorithm is part of the evaluation function to determine the goodness of the selected feature subset. Wrappers can usually achieve better results than filters while filters are more general and computationally less expensive than wrappers [7]. A FS algorithm explores the search space of different feature combinations to reduce the number of features and simultaneously optimize the classification performance. Support vector machines (SVM) are becoming increasingly popular in the machine learning and computer vision communities. Training a SVM requires the solution of a very large quadratic programming (QP) optimization problem. In this paper, we use a variant of SVM for fast

training using sequential minimal optimization (SMO) [23].

5. Conclusion and Future Aspect

This paper proposes an envisioning framework for intrusion detection i.e. Hybrid Intrusion Detection System. The developed framework is an intelligent, adaptive and effective intrusion detection framework. The experimental analysis is performed on the developed IDS framework and is compared with other techniques present in the scenario. The resultants obtained convey that the developed hybrid framework is highly effective to overcome the deficiencies found in previous work. As the framework uses two data mining techniques (i.e. TAN and SVM) to breed the classification rules, it can be effortlessly implemented in real time and is able to detect and adapt new types of intrusive activities. Also experimental assessment shows that the developed framework has reduced the false alarm rate and increased the accuracy up to noteworthy extend which is a major concern in case of intrusion detection mechanism. In addition to this, the framework is able to detect U2R and R2L attacks more efficiently than previous findings, boosting up the detection process. In future, some more work can be made in order to detect U2R and R2L attacks more accurately which may tend to further enhance the system efficiency.

Acknowledgment

It is not only customary but necessary for a researcher to mention her indebtedness to those who had helped in carrying out and enhance the research work. I pay my deep regards to God, my Parents and my loving Friends for their support and wishes which made this tedious work easy and successful. Finally, I would like to extend my thanks to all those who have contributed, directly or indirectly to make this project successful.

References

[1] KDDCUP-99 task description. https://kdd.ics.uci.edu/databases/kddcup99/task.html.

[2] Deepthy K Denatious & Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection", International Conference on Computer Communication and Informatics (ICCCI -2012), Jan. 10 – 12, 2012, Coimbatore, INDIA

[3] Srinivas Mukkamala, Andrew H. Sung, Ajith Abraham, Intrusion detection using an ensemble of intelligent paradigms, Elsevier, Journal of Network and Computer Applications 28 (2005) pp.-167–182.

[4] Sandhya Peddabachigari, Ajith Abraham,Crina Grosan, Johnson Thomas, Modeling intrusion detection system using hybrid intelligent systems, Elsevier, Journal of Network and Computer Applications 30 (2007), pp.114-132.

[5] KddCup99 dataset, available at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999.

[6] Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, "Using Rough Set And Support Vector Machine For Network Intrusion Detection", International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, April 2009

[7] Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324

[8] Wei-Hao Lin and Alexander Hauptmann, Meta-classification: Combining Multimodal Classifiers, Springer, Mining Multimedia and Complex Data, LNAI 2797 (2003) pp. 217–231.

[9] Peddabachigiri S., A. Abraham., C. Grosan and J. Thomas, "Modeling of Intrusion Detection System Using Hybrid intelligent systems" , Journals of network computer application, 2007

[10] Mrutyunjaya Panda and Manas Ranjan Patra, "A Comparative Study Of Data Mining Algorithms For Network Intrusion Detection", First International Conference on Emerging Trends in Engineering and Technology, pp 504-507, IEEE, 2008

[11] M.Govindarajan and Rlvl.Chandrasekaran, "Intrusion Detection Using k-Nearest Neighbor" pp 13-20, ICAC, IEEE, 2009

[12] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi and Lilly Suriani Affendey, "Intrusion Detection Using Data Mining Techniques", pp 200-203, IEEE, 2010FRNN(U, C, y)

[13] Roshan Chitrakar and Huang Chuanhe, "Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes Classification", IEEE,2012

[14] David Ndumiyana, Richard Gotora and Hilton Chikwiriro, "Data Mining Techniques in Intrusion Detection: Tightening Network Security", International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 5, May – 2013

[15] Muhammad K. Asif, Talha A. Khan,Talha A. Taj, Umar Naeem and Sufyan Yakoob, " Network Intrusion Detection and its Strategic Importance", Business Engineering and Industrial Applications Colloquium(BEIAC), IEEE, 2013

[16] Kapil Wankhade, Sadia Patka and Ravindra Thools, "An Efficient Approach for Intrusion Detection Using Data Mining Methods", IEEE 2013

[17] Vaishali B Kosamkar and Sangita S Chaudhari, "Data Mining Algorithms for Intrusion Detection System: An Overview", International Conference in Recent Trends in Information Technology and Computer Science (ICRTITCS), 2012

[18] Iwan Syarif, Adam Pruge Bennett and Gary Wills, "Unsupervised clustering approach for network anomaly detection", IEEE.

[19] Wei-Hao Lin and Alexander Hauptmann, Meta-classification: Combining Multimodal Classifiers, Springer, Mining Multimedia and Complex Data, LNAI 2797 (2003) pp. 217–231.

[20] Alexandra M. Carvalho, Arlindo L. Oliveira and Marie-France Sagot, Efficient learning of Bayesian network classifiers: An extension to the TAN classifier, Proceedings of Advances in Artificial Intelligence, Springer, Volume 4830, (2007), pp 16-25.

[21] Ajit Singh, Tree-augmented naive bayes, Homework 2 Problem 7 of Probabilistic Graphical Models, Fall 2006.

About The Author

Namita Parati is working as Assistant Professor at Bhoj Reddy Engineering College for Women, Hyderabad, INDIA. She has received B.E, M.Tech Degree in Computer Science and Engineering. Her main research interest includes intrusion detection using hybrid network.

Sumaltha Potteti is working as Assistant Professor at Bhoj Reddy Engineering College for Women, Hyderabad, INDIA. She has received B.Tech, M.Tech Degree in Computer Science and Engineering. Her main research interest includes Cloud computing and intrusion detection