Robust Ensemble Model for Intrusion Detection using Data Mining Techniques

Reshamlal Pradhan, Deepak Kumar Xaxa

Abstract— The current generations increasingly rely on the internet and advanced technologies. As network attacks have increased over the past few years, intrusion detection system (IDS) is increasingly becoming a critical component to secure the network. Due to large volumes of security audit data as well as complex and dynamic properties of intrusion behaviors, optimizing performance of IDS becomes an important open problem that is receiving more and more attention from the research community. In this paper we proposed Robust Ensemble Model for the classification of data using the data mining techniques to achieve high accuracy compare to individual models.

Index Terms— Classification, Confusion Matrix, Ensemble model, Feature Selection, Intrusion Detection System (IDS), Data Mining Techniques, Statistical Techniques.

1 Introduction

With rapid development of internet and intranet in today's life, Information/Network security is becoming an important issue for any organization to protect data and information in their computer network against various types of attack with the help of an efficient and robust Intrusion Detection System (IDS). Today Intrusion detection system (IDS) is a necessary addition to the security infrastructure of most organizations[1,2].

Intrusion detection systems (IDSs) are responsible for monitoring the events occurring in a computer system or network, analyzing them for signs of security problems (intrusions) defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer or network. Intrusions are caused by attackers accessing the systems from the Internet, authorized users of the systems who attempt to gain additional privileges for which they are not authorized, and authorized users who misuse the privileges given them.

IDS can be developed using various machine learning techniques like Classification, prediction etc. IDS is a classifier which classifies the data as normal or attack. Classification is one of the very common applications of the data mining in which similar type of samples are grouped together in supervised manner. A simple working of classifier is depicted in figure1.

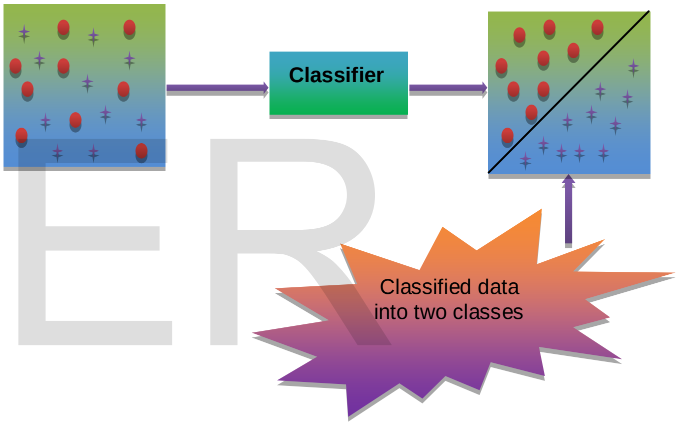

Fig 1: Classification of data into two classes

1.1 Types of Attacks

In today's scenario there are various kinds of attacks found in computer network[1,2]. Some of general attacks are:

DoS (Denial of Service) attacks- DoS attacks attempt to shut down a network, computer, or process, or otherwise deny the use of resources or services to the authorized users.

Probe (probing, scanning) attacks- Attacker uses network services to collect information about a host (e.g. list of valid IP addresses, what services it offers, what is the operating system).

Compromises- attackers use known vulnerabilities such as buffer overflows and weak security to gain privileged access to hosts.

R2L (Remote to Login) attacks - attacker who has the ability to send packets to a machine over a network (but does not have an account on that machine), gains access (either as a user or as a root) to the machine and does harmful operations.

U2R (User to Root) attacks - attacker who has access to a local account on a computer system is able to elevate his or her privileges by exploiting a bug in the operating system or a program that is installed on the system.

- *Reshamlal pradhan is currently pursuing masters degree program in computer science engineering in MATS University,Raipur(C.G.),INDIA PH-0771 4078994, 95, 96, 98. E-mail: reshamlalpradhan6602@gmail.com*
- *Mr. Deepak kumar xaxa is currently asst prof in computer science department in MATS University,Raipur(C.G.),INDIA, PH-0771 4078994, 95, 96, 98. E-mail: xaxadeepak@gmail.com*

Trojan horses / worms– attacks that are aggressively replicating on other hosts (worms – self-replicating; Trojan horses are downloaded by users).

1.2 Types of IDS

Intrusion detection systems (IDS)can be classified into different ways[1]:

Active and passive IDS:

An active Intrusion Detection Systems (IDS) is also known as Intrusion Detection and Prevention System (IDPS). Intrusion Detection and Prevention System (IDPS) is configured to automatically block suspected attacks without any intervention required by an operator. Intrusion Detection and Prevention System (IDPS) has the advantage of providing real-time corrective action in response to an attack.

A passive IDS is a system that's configured to only monitor and analyze network traffic activity and alert an operator to potential vulnerabilities and attacks. A passive IDS is not capable of performing any protective or corrective functions on its own.

Network Intrusion detection systems (NIDS) and Host Intrusion detection systems (HIDS):

Network Intrusion Detection Systems (NIDS) usually consists of a network appliance (or sensor) with a Network Interface Card (NIC) operating in promiscuous mode and a separate management interface. The IDS is placed along a network segment or boundary and monitors all traffic on that segment.

Host Intrusion Detection Systems (HIDS) and software applications (agents) installed on workstations which are to be monitored. A host Intrusion detection systems (HIDS) can only monitors the individual workstations on which the agents are installed and it cannot monitor the entire network. Host based IDS systems are used to monitor any intrusion attempts on critical servers.

Knowledge-based (Signature-based) IDS and behavior-based (Anomaly-based) IDS:

There are two primary approaches to analyzing events to detect attacks: Signature-based detection and anomaly detection.

Signature-based detection, in which the analysis targets something known to be "bad", is the technique used by most commercial systems. It's also known as misuse detection. Anomaly detection, in which the analysis looks for abnormal patterns of activity.

A knowledge-based (Signature-based) Intrusion Detection Systems (IDS) references a database of previous attack signatures and known system vulnerabilities.

A Behavior-based (Anomaly-based) Intrusion Detection Systems (IDS) references a baseline or learned pattern of normal system activity to identify active intrusion attempts.

1.3 Data Mining Techniques

There are various data mining techniques[3,5,6,10], Decision tree is so popular because the construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data.

Classification and Regression Technique (CART): CART is one of the popular methods of building decision tree in the machine learning community. CART builds a binary decision tree by splitting the record at each node, according to a function of a single attribute. CART uses the gini index for determining the best split. The initial split produces the nodes, each of which we now attempt to split in the same manner as the root node. Once again, we examine the entire input field to find the candidate splitters. If no split can be found then significantly decreases the diversity of a given node, we label it as a leaf node. Eventually, only leaf nodes remain and we have grown the full decision tree.

Chi-Squared Automation Interaction Detection (CHAID): CHAID is a derivative of AID (Automatic Interaction Detection), it attempts to stop growing the tree before over fitting occurs. CHAID avoids the pruning phase. In the standard manner, the decision tree is constructed by partition the data set into two or more data subsets, based on the values of one of the non-class attributes. After the data set is partitioned according to the chosen attributes, each subset is considered for further partitioning using the same algorithm.

Iterative Dichotomizer 3 (ID 3): ID3 (Iterative Dichotomizer 3) used for constructing the decision tree from data. In ID3, each node corresponds to a splitting attribute and each arc is a possible value of that attribute. At each node the splitting attribute is selected to the most informative among the attributes not it considered in the path from the root.

Artificial Neural Network: Neural networks can be used for descriptive and predictive data mining. ANN is known as best classifier and is able to mine huge amount of data for classification. A neural network is composed of a set of elementary computational units, called neurons, connected together through weighted connections. These units are organized in layers so that every neuron in a layer is exclusively connected to the neurons of the preceding layer and the subsequent layer. Every neuron, also called a node, represents an autonomous computational unit and receives inputs as a series of signals that dictate its activation. Following activation, every neuron produces an output signal.

1.4 Statistical Techniques

Statistical models[3,5,6] involving a latent structure often support clustering, classification, and other data mining tasks. Because of their ability to deal with minimal information and noisy labels in a systematic fashion, statistical models of this sort have recently gained popularity.

Bayesian Net: Bayesian Net is statistical classifiers which can predict class membership probabilities, such as the probability that a given tuple belong to a particular class. Let X is a data sample whose class label is unknown. Let H be some hypothesis, such as that the data sample X belongs to a specified class C. For classification problems, we determine P (H|X), the probability that the hypothesis H holds given the observed data sample X.P (H|X) is the posterior probability, or a posteriori probability, of H conditioned on X.

$P(H\backslash X)=[P(X\backslash H) P(H)]\backslash P(X)$

Support Vector Machine (SVM): Support vector machines (SVMs) are supervised learning methods that generate input-output mapping functions from a set of labeled training data.

The mapping function can be either a classification function (used to categorize the input data) or a regression function (used to estimation of the desired output).

1.5 Ensemble Techniques

An ensemble model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. Bagging and boosting[1] are techniques that use a combination of models. Each combines a series of k learned models (classifiers), M1, M2…Mk, with the aim of creating an improved composite model M. bagging and boosting can be used for classification.

Combining the decisions of different models[1,4,8] means amalgamating the various outputs into a single prediction. The simplest way to do this in the case of classification is to take a vote (perhaps a weighted vote), in the case of numeric prediction, it is to calculate the average (perhaps a weighted average). Bagging and boosting both adopt this approach but they derive the individual models in different ways. In bagging, the models receive equal weight, whereas in boosting, weighting is used to give more influence to the more successful ones, just as an executive might place different values on the advice of different experts depending on how experienced they are.

BAGGING:

Basic intuition- The more data, the better the performance (lower the variance), so how can we get ever more data out of a fixed training dataset?

Method: construct \novel" datasets through a combination of random sampling and replacement

1) Each training instance has a (1- 1 /N)N probability of being replaced by a random training instance.
2) Generate k permutations of the training dataset and build a base classifier over each.
3) Combine the base classifiers via voting.

BOOSTING:

Basic intuition- Tune base classifiers to focus on the “hard to classify" instances.

Method: Iteratively change the distribution and weights of training instances to reflect the performance of the classifier on the previous iteration

1) Start with each training instance having a 1/N probability of being included in the sample.
2) Over T iterations, train a classifier and update the weight of each instance according to its ability to classify training instances.
3) Combine the base classifiers via weighted voting.

1.6 Feature Selection

Feature selection[1,2] consists of detecting the relevant features and discarding the irrelevant ones, with the goal of obtaining a subset of features that describes properly the given problem with a minimum degradation of performance. In complex classification domains, Features may contain false correlations, which hinder the process of detecting intrusions. Further, some features may be redundant since the information they add is contained in other features. Extra features can increase computation time, and can impact the accuracy of IDS. Feature selection improves classification by searching for the subset of features, which best classifies the training data. Different feature selection techniques are CFS, IG, GR, and FVBRM.

Information Gain (IG):

This is one of the simplest (and fastest) attribute ranking methods and is often used in text categorization applications where the sheer dimensionality of the data precludes more sophisticated attribute selection techniques. If A is an attribute and C is the class, following equations given the entropy of the class before and after observing the attribute.

$H(C) = -\Sigma p(c) \log2(c)$,

$H(C|A) = -\Sigma P(a) \Sigma P(c|a) \log2P(c|a)$

The amount by which the entropy of the class decreases reflects the additional information about the class provided by the attribute and is called information gain. Each attributes Ai itself and the class:

$IGi = H(C) - H(C|Ai) = H(Ai) - H(Ai|C) = H(Ai) + H(C) - H(AiC)$

Gain Ratio (GR):

The information gain measures prefer to select attributes having a large number of values. The gain ratio an extension of info gain, attempts to overcome this bias. Gain ratio applies normalization to info gain.

Correlation Based Feature Selection (CFS):

Correlation based Feature Selection (CFS) is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features.

2 RELATED WORK

The field of intrusion detection and network security has been around since late 1980s. Since then, a number of methods and frameworks have been proposed and many systems have been built to detect intrusions. Various techniques such as association rules, clustering, naïve Bayes classifier, support vector machines, genetic algorithms, artificial neural networks, and others have been applied to detect intrusions. Some of recent work on intrusion detection is-

Nagle et.al [1] have proposed Feature Extraction Based Classification Technique for IDS. They have worked on naïve bayes and J48 classification techniques on NSL KDD dataset. They also have stated that combination of decision tree with statistical techniques would be result in high accuracy.

Mukherjee et.al [2] have proposed ”Intrusion detection using bayes classifier with feature reduction”. They have worked on naïve bayes classifier on NSL KDD dataset and used a new

feature selection technique FVBRM to reduce the no of features and perform intrusion detection using reduced feature subsets. Proposed feature selection technique FVBRM provided classification accuracy 97.78% with 24 features.
Y. Li et.al [9] have proposed ”An efficient intrusion detection system based on support vector machines and gradually feature removal method”, and have worked on support vector machine classification technique on KDD 99 dataset and have used feature selection. SVM gives highest accuracy 98.62% in case of gradually feature reduction (GFR) with 19 features.
L. Koc[7] have proposed ”A network intrusion detection system based on hidden naive byes classifier”. They have worked on hidden naïve bayes classifier on KDD99 dataset. Proposed model HNB with proposonal k-interval discretization and INTERACT feature selection method achieve 93.72% which is higher than traditional naive bayes.

3 Proposed Framework

Here we have proposed a robust ensemble intrusion detection model. Classification model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute.
Classification consists of two steps:

- Training: It is the supervised learning of a training set of data to build a model.
- Testing: It includes classifying the data according to that model.

The model consists of two phase- model building and model validation. Model building performs on training set of data and Model validation performs on test set of data. The framework is given in figure2.

Model Building Phase
First step perform in model building phase is preprocessing of training set of data. Preprocessing consist of normalization to remove noise and getting data linearity. Database of frequent itemsets for attack-free data is made in static level mining. For entire training data, find suspicious frequent itemsets in dynamic level mining that are not in the “attack-free” database. Train a classifier to classify itemset as known attack, unknown attack or normal event through reduced feature subset.

Model Validation Phase
In this phase also first preprocessing of training data set is performed. Preprocessing provides data linearity and noise removal of test data set. Dynamic mining module produces suspicious itemsets from test data. Along with features from feature selection module, itemsets are fed to classifier and the classifier classifies the itemset as attack or normal.

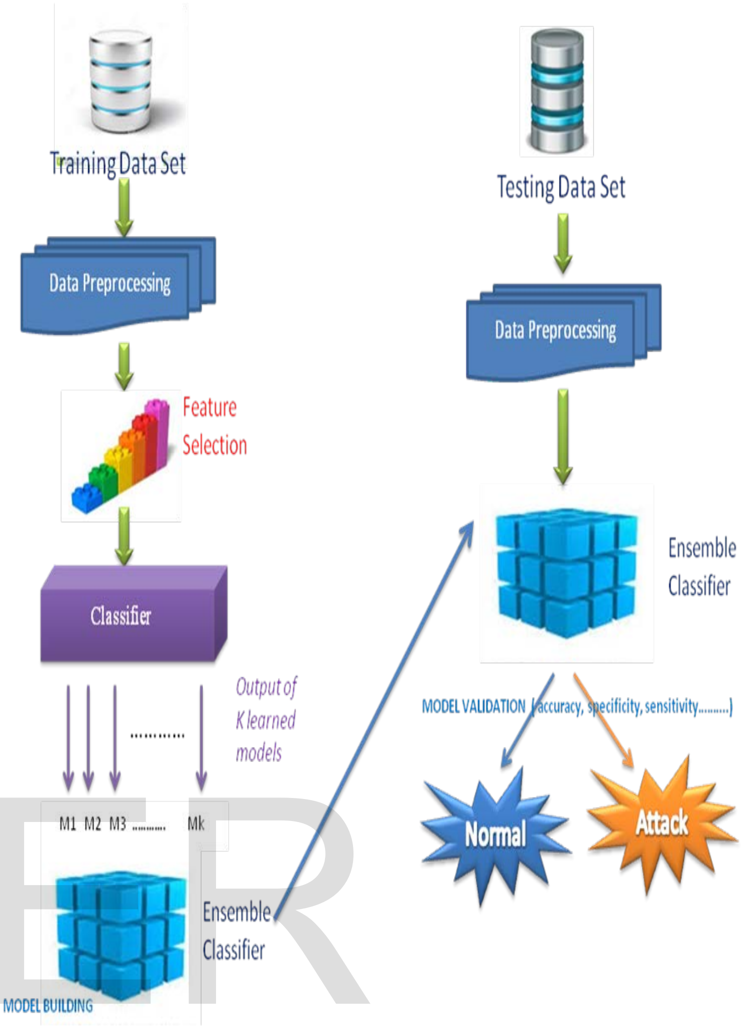

Fig 2: Ensemble intrusion detection model.

Classifier
A classifier first finds suspicious itemset through static and dynamic mining from preprocessed dataset. Then label the suspicious itemset as normal or attack.The framework given in figure 3.

Ensemble Classifier
Ensemble classifier is a combination of two or more classifier to provide the better performance than individual classifiers. Here during model building phase first we perform classification through sever classifiers and then Combine the decisions of different models. In other words amalgamate the various outputs into a single prediction through bagging or boosting ensemble technique[1].
During Model validation phase we fed the test dataset into ensemble classifier which classify the data as normal or attack with high accuracy.

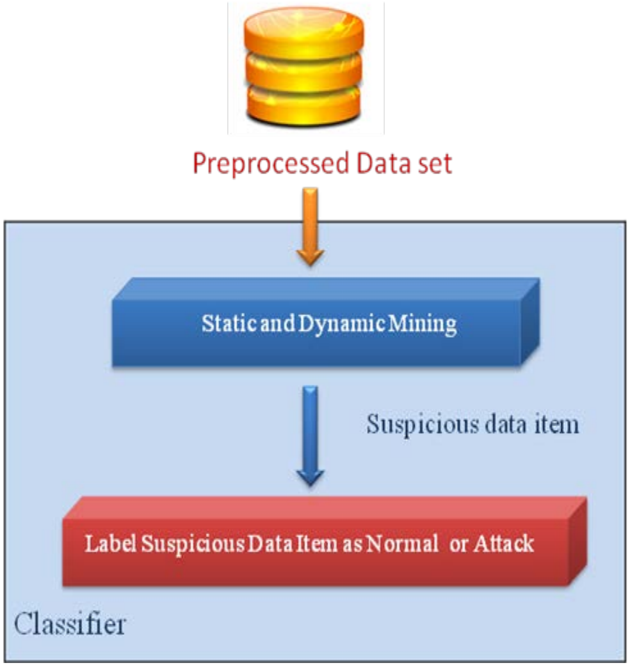
Preprocessed Data set
Static and Dynamic Mining
Suspicious data item
Label Suspicious Data Item as Normal or Attack
Classifier

Fig 3: Classifier

4 Model Evaluation Criteria

Classifier result is evaluated using confusion matrix[1].
Confusion Matrix- This may be used to summarize the predictive performance of a classifier on test data.
It is commonly encountered in a two-class format, but can be generated for any number of classes. A single prediction by a classifier can have four outcomes which are displayed in the following confusion matrix. Table 1 depicts confusion metrics.

TABLE 1
Confusion Metrics

		PREDICTED	
		NEGATIVE	*PISITIVE*
ACTU-AL	*NEGATIVE*	*TN*	*FP*
	POSITIVE	*FN*	*TP*

The entries in the confusion matrix have the following meaning in the context of our study:
TN is the number of correct predictions that an instance is negative.
FP is the number of incorrect predictions that an instance is positive.
FN is the number of incorrect of predictions that an instance negative and
TP is the number of correct predictions that an instance is positive.
During testing phase, testing dataset is given as an input to the proposed technique and the obtained result is estimated with the evaluation metrics namely precision, recall and Accuracy.
The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the equation: AC = (TP+TN) / (TP+FN+FP+TN).
The recall or sensitivity or true positive rate (TPR) is the proportion of positive cases that were correctly identified, as calculated using the equation: TPR = TP / (TP+FN).
The false positive rate (FPR) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation. FPR = FP / (TN+FP).
The true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation: TNR = TN / (TN+FP).
The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation: FNR = FN / (TP+FN).
The precision (P) is the proportion of the predicted positive cases that were correct, as calculated using the equation: Precision = TP / (TP+FP).

5 Conclusion

This paper presented a comprehensive taxonomy and state of the art of intrusion detection and prevention systems to drag researchers' attention for possible solutions to intrusion detection and prevention. The proposed model has shown in figure in which data set is divided into two partitions as training and testing. Training data will use for model building and testing data for model validation. The proposed model with ensemble technique using reduced feature subset will be result in high accuracy compare to each individual model.

As future work, we plan to ensemble different combinations (data mining techniques) to maintain the good results of the minority classes but enhancing those of the majority ones.

References

[1] Manish Kumar Nagle1, Dr. Setu Kumar Chaturvedi- Feature Extraction Based Classification Technique for Intrusion Detection System, International Journal of Engineering Research and Development (August 2013).

[2] Dr. saurbh Mukherjee (2012) ,"Intrusion detection using bayes classifier with feature reduction", Procedia technology".

[3] Jiawei Han, Micheline Kamber, (2006), "Data mining concepts and techniques", Second edition, San Francisco, Margan Kaufmann Publishers, USA,.

[4] Amuthan Prabakar Muniyandi, R. Rajeswari, R. Rajaram - Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm, International Conference on Communication Technology and System Design 2011.

[5] Arun K. Pujari. (2001). Data mining techniques, 4th edition, Universities Press (India) Private Limited.

[6] David L. Olson, Dr. Dursun Delen ”Advanced Data Mining Techniques “, NewYork, 2008.

[7] L. Koc (2012) ”A network intrusion detection system based on hidden naive byes multiclass classifier”, Journal of Expert system with publications”.

[8] Mrutyunjaya Panda, (2011)”A hybrid intelligent approach for network intrusion detection”, Proceedia Engineering.

[9] Y. Li (2012),”An efficient intrusion detection system based on support vector machines and gradually feature removal method”, Expert systems with Applications”.

[10] Krzysztof Cios, et al.(2000), “ Data mining methods for knowledge discovery”, Third Edition, Kluwer academic publishers.