I. J. Computer Network and Information Security, 2012, 8, 8-20
Published Online August 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2012.08.02

Cascading of C4.5 Decision Tree and Support Vector Machine for Rule Based Intrusion Detection System

Jashan Koshal, Monark Bag
Indian Institute of Information Technology Allahabad, Uttar Pradesh-211012, India
jashankoshal@gmail.com, monarkbag@gmail.com

Abstract **—** Main reason for the attack being introduced to the system is because of popularity of the internet. Information security has now become a vital subject. Hence, there is an immediate need to recognize and detect the attacks. Intrusion Detection is defined as a method of diagnosing the attack and the sign of malicious activity in a computer network by evaluating the system continuously. The software that performs such task can be defined as Intrusion Detection Systems (IDS). System developed with the individual algorithms like classification, neural networks, clustering etc. gives good detection rate and less false alarm rate. Recent studies show that the cascading of multiple algorithm yields much better performance than the system developed with the single algorithm. Intrusion detection systems that uses single algorithm, the accuracy and detection rate were not up to mark. Rise in the false alarm rate was also encountered. Cascading of algorithm is performed to solve this problem. This paper represents two hybrid algorithms for developing the intrusion detection system. C4.5 decision tree and Support Vector Machine (SVM) are combined to maximize the accuracy, which is the advantage of C4.5 and diminish the wrong alarm rate which is the advantage of SVM. Results show the increase in the accuracy and detection rate and less false alarm rate.

Index Terms **—** Intrusion Detection System, Data Mining, Decision Tree, Support Vector Machine, Hybrid Algorithm

I. Introduction

Heady et al 1990 defined intrusion as "any set of activities that tries to compromise the reliability, confidentially or accessibility of a resource" [1]. Damaging the information or accessing it as an unauthorized user, or maliciously using the information comes under the category of intrusion. Six types of attacks are defined by Mahoney [2] and they are viruses, worms, server attacks, client attacks, network attacks and root attacks.

It is difficult to detect such kinds of attacks despite of strong security policies, anti- virus software, firewalls or other mechanism because every system has some weakness and bugs. That's why the IDS are premeditated and can detect the new attacks.

IDS monitor all traffic in a network, and identify the suspicious, malicious activities. It also continuously monitors the system health and responds to the supervisor if anything goes wrong. The success of IDS is that it must be intelligent to diagnose all the attacks in the Local Area Network (LAN).

IDS should able to discriminate among the normal and attacks and should take the balance between the false negative and alarm rate.

In today's scenario network is a heart of the communication. Lot of things individual can do on internet. Internet gives many advantages and but has some disadvantages too. It is used as tool for crime. One of the major and famous crimes is hacking. In this attack, one can gain the information of other system by accessing it as a root user and can use this data for profit purpose. In today's arcade there are lots of computer peripherals available and the internet technology has got a tremendous amount of growth, moving of data from one domicile to another is a very critical issue in this present world. Internet plays a leading role in information transmission. But internet is not fully secured. There are many threats to internet. So there is a severe need to make them secure and safe.

Network security has become a critical issue owing to the incredible growth of computer networks usage. It is technically hard and economically expensive for the manufactures to secure the computer systems from external attacks. New application areas have been recognized in the field of computer networks in past twenty years due to the swift progress in Internet based technology. Local Area Network (LAN) and Wide Area Network (WAN) are increasingly being used in business, financial, industry, and security and healthcare sectors, due to their great progress. And this resulted in making us further more dependent on the computer networks. These application areas are making computer networks a target for abuse and a big weakness for the community.

Prevailing tensions of hacking are enhanced by new entities like worms, viruses Trojans. Our existing defense mechanisms in the network are weak. We should realize that the threat could have disastrous after effects when we consider the high popularity, connectivity and the ever-

growing dependency on them. Thus, the research in this concern should be given a very great priority.

There has been a massive evolution in the figure of attacks, so to secure the information, intrusion detection is being applied. Firewalls do deliver some defense, but do not provide full protection so it desires to be used along with an intrusion detection system. IDS are obligatory to deal with such kind of attacks and cases. Intrusion detection provides help to computer systems to handle the attacks. Intrusion detection system collects information from various available sources within the computer system and networks. This statistics is matched with the already defined attack patterns and detects the attacks and vulnerabilities.

Intrusion detection systems are being developed as devices to detect attacks and anomalies in the network, and thus are becoming very important. IDS are useful in detecting successful intrusion, and also in monitoring the network traffic and the attempts to break the security. Intrusion detection is the practice of observing and examining the actions going on in a system in order to identify the attacks and susceptibilities.

There are some terminologies involved in this research work and it is represented in Table 1 below.

Table 1 Terms and their Definition

Terms	Definition
Alert/Alarm	It a signal that is generated by the system to report the administrator that the system has been or is being attacked.
True Positive (TP)	Attack detected by the system and the signal is raised.
False Positive (FP)	System generates the alarm when there is no attack detected.
False Negative (FN)	A failure of IDS to detect an actual attack.
True Negative (TN)	No attack is identified and no alarm is raised.
Noise	Unwanted data that causes the system to raise the alarm.
Alarm Filtering	Process to distinguish between false positives and actual attacks.
SVM	Support Vector Machine
IDS	Intrusion Detection System

Categories of intrusion detection system are Host-Based and Network-Based Intrusion Detection System, classified according to different data source and Misuse and Anomaly Intrusion Detection System, classified according to different analysis method.

a) Host Based IDS- For this type of system the data source is the records of activities done by host, log of the operating systems, logs of application etc.

Advantage

1. Subsequently in this kind of system the data comes from the system log itself, so it can detect more accurately whether the host is intruded or not than network based IDS.
2. No installation of additional hardware. Specific software deployment is required on the host.

Disadvantage

1. As the software is required to be installing on each specific host, the cost is very high. Meanwhile the logs for every system is different, every system requires IDS to be installed.
2. Efficiency of the hosts is affected as system is installed on every system and it may require system sources.

b) Network-based IDS- Its data is mainly collected network generic stream going through network segments, such as: Internet packets. And its advantage and disadvantage are stated as follows:

Advantage

1. Low installation cost as it can detect all the attacks in the LAN.
2. Host based IDS cannot detect some types of attack such as DOS etc. These sorts of attacks can be identified by network based IDS.

Disadvantage

1. Due to network congestion, there is a loss of packets in a network, so network based IDS cannot detect all packets.
2. If network size is large, the system will require more CPU and resources to examine the packets.
3. Network based IDS cannot detect encrypted packets.

c) Misuse Detection- Signature based detection is also the name of misuse detection. There is a database of past attack signature. To detect whether the current signature is an attack or not is compared with the signature existent in the database. If comparisons is found, then it is an attack, if not than it is a normal signature. Signature based IDS has better detection rate and false alarm rate for known attacks, but it has low detection rate when it comes to unknown attacks. There is another problem that the database should be kept restructured from time to time regularly.

d) Passive Intrusion Detection- This kind of system generates an alarm/alert when an attack is detected and sends it to the administrator. Now it is up to the supervisor how to deal with the issue, whether to block the action or reply in some other way.

e) Reactive Intrusion Detection- This system will alert the administrator with the alarm and also block the network traffic from that source by taking some action or reprogramming the firewall.

In this paper we have proposed network-based IDS by combining two data mining algorithms C4.5 Decision Tree and SVM. The main focus is to combine the advantages of both the algorithms in order to reduce the false alarm rate and to increase the accuracy and detection rate. The remainder of the paper is arranged as follows. Section II describes about the past work that been done on this area. This section will check into the advantages and disadvantages of the various other approaches that have been used in developing the system. Methodologies involved in implementing the system have

been described in detail in section III. Results obtained by the implementation of the algorithms have been discussed in section IV. In the section V, conclusions were drawn and future works have been proposed.

II. RELATED WORK

The approaches that have been presented to develop the system by the researchers have been discussed. First will see the single methods that have been applied, then we go through the hybrid methods applied in making the intrusion detection system.

There are various methodologies that have been applied in the creation of IDS and mainly it is classified as single and hybrid approaches.

A) Single Approaches

Classification Algorithm- Classification is the process of discovering or finding a model that elaborates and categorizes the data classes or concepts from the databases. The data mining systems can also be measured on the basis of "Which databases are mined, which type of knowledge are mined, which type of techniques are utilized, which applications are adapted?"

Steven L. Scott (2004) [3] defines a model based methodology for the constructing of intrusion detection system and considers the general methods that are proficient of being applied to different networks using specific algorithms. With the help of Bayesian methods, hierarchical models are built which lead to the development of coherent systems that can handle complex distribution present in the network.

Giorgio Giacinto et al (2003) [4] proposed a pattern recognition methodology based on the union of multiple classifiers for the network intrusion detection. In all five decision fusion methods are deployed in the experiment and their performance are compared. Classifier fusion is employed and its performance and other parameters are evaluated and discussed. The reported results showed that the MCS approach provides a better trade-off between generalization abilities and false alarm generation than that provided by an individual classifier trained on the overall feature set.

To solve the multi-class classification problem, Gang Kou et al (2009) [5] re-examine the MCLP and MCNP models and after that built a multi-class MCMP model. It was at that point tested on two network intrusion datasets. They demonstrate that the model so proposed can achieve low false alarm rate and have high classification accuracies.

Inho Kang, M. K. Jeong (2011) [6] suggest a new one-class classification method to raise the performance of IDS. Centred on three diverse points of view such as contents, dimension and structure they also propose some new mined features for host based intrusion detection.

Clustering Techniques- Clustering is the practice of combining the records into classes or clusters, so that entities with in the cluster have high resemblance in compare to one another on the further hand are very dissimilar to entities in other clusters. There are alternate approaches which are used in clustering process: Partitioning methods, Hierarchical methods, etc.

Sheng Yi Jiang et al (2006) [7] suggest a novel method to calculate the cluster radius threshold and an improved nearest neighbour (INN) method for data classification. They proposed clustering based unsupervised intrusion detection (CBUID) whose linearity is proportional to the magnitude of dataset and the quantity of attributes. Their method surpasses the existing method and achieves detection rate that was high and false alarm rate that was low.

Seungmin Lee et al (2011) [8] employ K means clustering with SOM so that the model developed becomes self-adaptive and dynamic in nature. Experiment were carried out on well-known data set KDD cup 99, and results shows that approach can growth the detection rate while making the false alarm rate low and also proficient of identifying new types of attacks.

V. Nikulin (2006) [9] applied the concept of threshold based clustering. Main motive of applying the clustering is to shrink the number of signatures, and to reduce as such much as possible the number of comparison required to categorize the new input. Experiments were carried out on KDD cup 99 and the results shows that procedures were effective.

Fuzzy Logic- Fuzzy Logic is a problem-solving control structure approach that gives itself to implementation in the systems which are ranging from multi-channel PC or workstation acquisition and control systems. It can be engaged in hardware, software, or in both. It offers a simple manner to attain on a definite decision based upon indefinite, ambiguous, inaccurate, noisy, or absent input information.

The core of IDS, the classification engine uses Association Based Classification. Fuzzy association rule are employed by Arman Tajbakhsh et al (2009) [10] for the building the classifier. The similarity between any new sample with different class of rule set are analyzed by using matching measure and the class corresponding to the rule set that is matched accurately is termed as label of the said sample. A new methodology to hustle up the rule induction procedure via decreasing items is proposed that may be involved in extracted rules. Dataset that is used to evaluate doesn't show that promising results but the false positive rate is minor while overall detection rate and detection rate of well-known attacks are significant.

Dickerson and Dickerson (2000) [11] developed a Fuzzy Intrusion Recognition Engine which is an anomaly based IDS. It uses fuzzy logic for the recognizing the activity as malicious. It takes the help of simple data mining techniques for the processing the network data. Components of FIRE are explained and by what means data mining can help in this purpose. Results of test that was applied on the network data shows that FIRE can detect common attack types.

Genetic Algorithms- Genetic Algorithms are adaptive exploratory search algorithms which are introduced on the evolutionary concepts of normal selection and genetic. This heuristic is normally used to create valuable

solutions to optimization and search problems. With the help of inheritance, crossover, selection etc. generate solutions to optimize the problem, thus genetic algorithm belongs to evolutionary algorithm.

Intrusion Detection Based on Genetic Clustering (IDBGC) algorithm is proposed by Y. Liu et al (2004) [12]. Clusters are established automatically and intruders are detected by labelling them under normal or abnormal groups. Algorithm was simulated and it proved to be the effective for intrusion detection.

Adel Nadjaran Toosi (2007) [13] incorporates soft computing paradigms like neuro-fuzzy networks, fuzzy inference approach. Initial classification is done by neuro- fuzzy classifiers and then the based on the result of neuro-fuzzy classifier, fuzzy inference system make the final decision whether the activity is normal or abnormal.

Kamran Shafi et al (2009) [14] present a supervised learning classifier system that dynamically and adaptively learn signatures for intrusion detection. Signatures are discovered by the classifier and are added to the knowledge base. Their approach is a hybrid that learns both intrusive and normal behaviour. Performance is evaluated with the publically available dataset for intrusion detection and results shows that the offered system is effective.

Three kinds of genetic fuzzy systems are proposed based on Michigan, Pittsburgh and Iterative Rule Learning (IRL) approaches by Mohammad Saniee Abadeh et al (2011) [15]. Some results are showed and the compares the performance of three genetic fuzzy system.

Chi-Ho Tsang et al (2007) [16] Multi-objective genetic fuzzy intrusion detection system (MOGFIDS) is proposed. Agent based evolution framework is employed to extract precise and interpreTable fuzzy rule based knowledge for classification.

Machine Learning- Machine learning is a technique of predicting the properties, focuses on the unknown attributes of the data, and the system deploying machine learning algorithm are capable of taking definite action accordingly.

Using a supervised machine learning technique, PhurivitSangkatsanee et al (2011) [17] proposed a real time intrusion detection system. Several techniques were applied for the development and results show that decision tree outperforms the other approaches. Then using decision tree algorithm they developed the real time intrusion detection system. For increasing the reliability and detection precision, and to decrease the false alarm rate, they developed a new post- processing procedure.

Yang Yi et al (2011) [18] investigated the incremental training algorithms of the network intrusion detection, and proposed an improved incremental SVM algorithm. The SVM algorithm is combined with U-RBF, the modified kernel function, to network intrusion detection. Computer simulations demonstrate that the suggested algorithm simplifies the oscillation phenomenon in the incremental learning process and saves training and prediction time.

Eleazar Eskin [19] proposed a new geometric structure for unsupervised anomaly detection. The data elements are plotted to a feature space and anomaly detection is done by determining which points lay in regions of feature space. Two feature maps are proposed that are data- dependent normalization feature map and spectrum kernel. Three algorithms are employed for detecting the points in the feature space. Their algorithms were able to detect attack from unlabelled data.

Neural Networks- A neural network comprises of nodes and edges. The assessment of the weight on edge defines how a node affects adjacent node. There are two subsets of the nodes one is the input nodes and other is the output nodes. Neural networks have several advantages like inherently parallel, distributed architectures; learn by adjusting weights, attributes extractor, etc.

A. K. Ghosh et al (1999) [20] Three anomaly detection practices are offered in this paper for profiling program behaviour that evolves from memorization to generalization. The aim of monitoring behavior of program is to enable detection of potential intrusions by keeping track of irregularities that occur in program behavior. The aim is to use machine learning procedures that can generalize from past observed behavior. DARPA dataset is used for estimating the performance of the system.

Chunlin Zhang et al (2005) [21] identify the suiTable method that can reduce the training time, provide great detection rate and low false alarm rate. Because of several advantages of neural networks, they applied Radial Basis Function, and results illustrates that this process has good performance for misuse and anomaly detection. Another objective was to develop IDS that can recognise both misuse and anomaly attacks and can adaptively train the module. A serial hierarchical IDS and parallel hierarchical IDS are the two frameworks proposed.

Jian Li (2004) [22] describes an IIDS based on ANN for anomaly detection. The structure takes network traffic data as input to analyse and categorize the behaviours of the legal users and detect the likely attacks. System has been verified and accepTable results have been obtained.

Mohammed Theeb Alotaibi [23] built a special IDS using Neural Network to detect the U2R attacks. System was labelled as a U2R Intelligent Detector (U2RID). Dataset used in the research was DARAPA. Research concluded that selecting the features with intrinsic information to train the Neural Networks can enhance the capabilities of the U2RID to detect U2R attacks.

Dewan Md. Farid et al (2010) [24] presented anomaly based NIDS using decision tree algorithm. Dataset used was KKD 99 and the achieved detection rate was 98% in comparisons with the other existing methods.

Statistical Techniques- Statistical techniques are also called as ‘top down learning’ approach. There are three different classes in statistical approach that are linear, nonlinear, decision tree.

W. Lee (2000) [25] describes the common intrusion detection framework (CIDF) to detect the intrusion in the

distributed environment. In this model, multiple IDSs can interchange information with each other. The system incorporate an ID model builder, which compute a new detection model with the help of data mining engine that receive audit data from novel attack from IDS. This new detection model is distributed to other IDSs.

Association Rule Mining- L. Hanguang et al (2012) [26] increases performance of the structure by applying the rule base deduced from Apriori algorithm, which is the standard of the association rule mining.

B) Hybrid Approaches

In this approach the two algorithms are combined for the development of IDS. Hybrid approach gives greatly improved results as equated to the single approaches. Various single approaches are pooled to form a hybrid algorithm for the development of IDS.

Gang Wang et al (2010) [27] concluded that ANN can deliver significantly improved performance of IDS compared with some traditional methods. A new approach called FC-ANN is proposed, based on fuzzy clustering and artificial neural network to expand the performance of IDS in terms of achieving high detection rate and less false alarm rate. Fuzzy clustering is applied first to generate different training subsets and then on this different training subset the different ANN models are accomplished to formulate different base model. At last fuzzy aggregation module is employed to combine these results.

Fusion of hierarchical clustering and support vector machines is suggested by Shi-Jinn Horng et al (2011) [28]. Hierarchical clustering provides the high qualified training instance to SVM reduces the training time and improve the performance of resultant SVM. Feature selection procedure was also applied to eliminate redundant features from the training set so that SVM model could classify the network data accurately. Overall performance was evaluated and is found to be worthy on comparing it with the other IDS.

Yinhui Li et al (2011) [29] introduced an IDS based on series of machine learning strategies, which has a following steps – compact data set is created by clustering the redundant data; apply the method ACO for selecting a proper small training data set; feature dimension are reduced from 41 to 19 so as to seize the key feature of the network; obtain the classifier with SVM and undertake a thorough prediction to the total KDD cup data set.

Ozgur Depren et al (2005) [30] suggested a novel intrusion detection system architecture using the anomaly and misuse detection approaches. There were 3 modules in this hybrid system that are misuse, anomaly detection module respectively, and a decision support system for conjoining the results of the two modules. SOM structure is used for anomaly detection and J.48 decision tree procedure is used to classify various kinds of attack.

To shrink the number of false positives, orthogonal and complementary approaches are presented by Tadeusz Pietraszek, Axel Tanner (2005) [31]. They use alert post processing with the support of conjoining data mining and machine learning. This method has been verified in various data sets, and it resulted in significant lessening in the aggregate of false positives in both simulated and real time environment.

Su-Yun Wu (2009) [32] compared two machine learning approaches in intrusion detection in terms of efficiency, including the classification tree and support vector machine, and provides a reference for the developing the intrusion detection system in future.

Tansel O ̈zyer(2007) [33] provide an intelligent intrusion detection system that uses two of the furthermost standard data mining algorithm, namely classification and association rules mining composed for guessing dissimilar behaviors in networked computers. They proposed a technique based on iterative rule learning via a fuzzy rule-based genetic classifier.

Sandhya Peddabachigaria et al (2005) [34] deploy Decision Tree, SVM for developing the IDS. After that they design a hybrid of these two models as DT-SVM model and further they proposed an ensemble approach with DT, SVM and DT-SVM models as base classifier. Ensemble approach gives the best performance and they showed that if proper base classifier is chosen then the accuracy can be 100%.

Combination of two anomaly based IDS that are Packet Header Anomaly Detection (PHAD) and Network Traffic Anomaly Detection (NETAD) with a misuse based IDS SNORT which is an open source software are presented by M. Ali Aydın et al (2009)[35]. Using the MIT Lincoln Laboratories network traffic records, the hybrid system is being evaluated and it result shows that the fusion IDS is a dominant system.

Siva S. SivathaSindhu et al (2011) [36] applied Wrapper based feature selection algorithm for the development of lightweight IDS. This algorithm maximizes the specificity and sensitivity of IDS. Neural ensemble decision tree iterative procedures are employed to evolve the optimal feature.

Shilpa lakhina et al (2010) [37] employed a new algorithm called Principal Component Analysis Neural Network Algorithm (PCANNA) to shrink the number of computer resources like memory, CPU time, to detect the attack. Neural network is used to identify the new attacks. NSL-KDD dataset is used for test and comparison. The proposed approach increases the classification accurateness and diminishes the number of input feature and time.

Z. Pan et al (2003) [38] consider the KDD99 dataset and proposed a hybrid system by applying BPNN and C4.5 Decision Tree. They also compare the performance of system by developing it with BPNN and not combining it with the C4.5 Decision Tree and discovered that the system cannot able to detect the User to Root (U2R) and Root to Local (R2L) network attacks.

Chittur (2001) [39] applied genetic algorithm and use a decision tree to characterize the data. To distinguish among the data, "the detection rate minus the false positive rate" is used as their preference criterion.

Ming-Yang Su (2011) [40] developed an anomaly based IDS using genetic algorithms and KNN (k-nearest neighbour) for the features selection and weighting. From

total of 35 features only 19 features were considered and accuracy of 97.42% was obtained and accuracy of 78% was recorded when 28 features were considered.

M. Jiang et al (2011) [41] represent a combined model for misuse and anomaly intrusion detection. Normal behaviour rule set are developed by using the clustering analysis algorithm to detect the new unknown attacks and association rule mining algorithm was useful to detect the known attack rapidly.

M. Panda, et al (2011) [42] proposed a hybrid intelligent intrusion detection system by combining the two classification algorithm for making the decision more accurate and rapid. First the classification or clustering was applied in the whole dataset and the resulted output is applied to another classification algorithm. They applied 10-fold cross validation method, and the result achieved is in the form of normal or intrusion.

A. Muniyandi et al (2011) [43] deployed K-means clustering and C4.5 decision tree procedure for detecting the intrusion in the internet environment. K cluster are made by partitioning the training dataset using K-means clustering and then on each cluster the decision tree was constructed. Decision tree on each cluster was exploited for the result.

Fig.1 shows the percentage wise distribution of the research paper under various methodologies that are applied in the creations of IDS. The most commonly and widely applied approach is the hybrid approach.

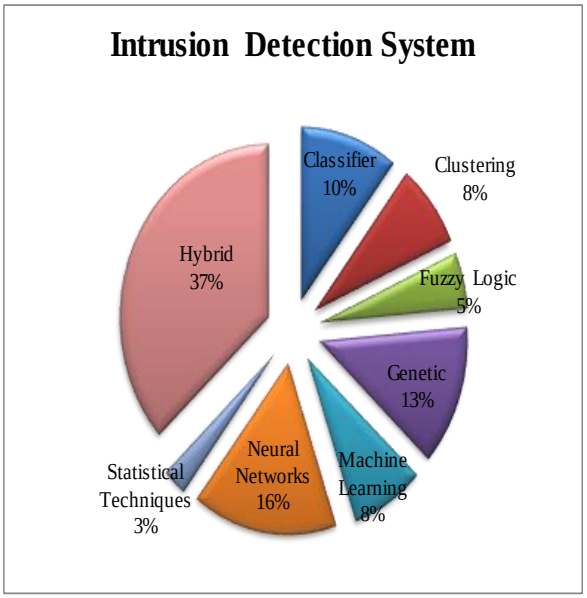


Fig 1 the percentage distribution of the number of papers under various IDS approaches

Hybrid approaches improves the accuracy of the IDS when compared to single approaches. Results from the different individual systems are combined to provide more accuracy and reliability. Researchers are focusing on hybrid methodology for developing the IDS as it can combine the advantages of two algorithms.

III. Implementation

Implementation is done using the tool WEKA which stands for Waikato Environment for Knowledge Analysis [44], implement in Java having java libraries which comprises of different data mining and machine learning algorithm. As it is built in java, it allows user to put on data mining plus machine learning algorithm to their data regardless of the platform and policy of the computer. It is liberally obtainable on internet and comes under GNU license.

This tool has the ability of preprocessing the data making it to be used by different algorithms, can analyze the performance of different classifier.

Classifiers are the main primary learning methods in WEKA. They produce the rule set or decision trees that facsimile the data. WEKA is specific of the simplest tools to bring out the technology in open environment.

A) Work Structure

Fig.2 shows the model of the intrusion detection system.

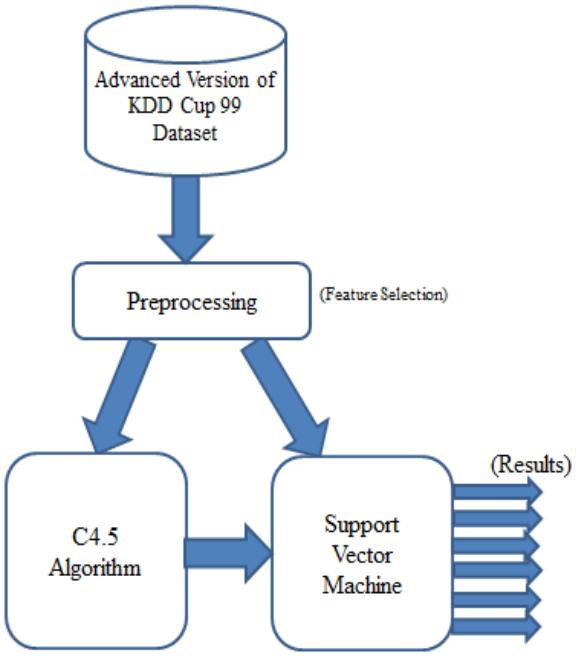


Fig 2 Hybrid C4.5 and SVM Model

B) Experiment Data Set

In this research the dataset that has been used is NSL-KDD, which is the advanced version of KDD Cup 99 for intrusion detection. There were some shortcomings in the KDD Cup 99 which affects the productivity of the system. This dataset consist of selected records of KDD Cup 99 and does not undergo from any of the limitations.

Shortcoming of the previous data set was that about 75%-78% records in both train data and test data are same respectively. The KDD Cup 99 data set contains redundant records. The algorithm becomes biased towards the frequent records because of the redundant records in the training set. The training model does not

able to learn the unfrequented records and therefore performance of the system declines.

The redundant records in train and test of KDD Cup 99 Dataset [45] are shown below in Table 2 and Table 3 respectively:-

Table 2: Statistics of redundant records in the KDD Train Set

	Original Records	Distinct Records	Reduction Rate
Attacks	3,925,650	262,178	93.32%
Normal	972,781	812,814	16.44%
Total	4,898,431	1,074,992	78.05%

Table 3: Statistics of redundant records in the KDD Test Set

	Original Records	Distinct Records	Reduction Rate
Attacks	250,436	29,378	88.26%
Normal	60,591	47,911	20.92%
Total	311,027	77,289	75.15%

C) Improvements to the KDD'99 Dataset

There are no redundant records in the train and testing dataset. Classifier will not be biased towards the frequent records. No need to randomly select the portion of data set for training and testing as both are equiTable. Evaluation of different research will be consistent in nature and comparable.

D) Feature Selection

There are two algorithms that come under feature selection and they are: - Correlation- Based Feature Selection (CFS) and Consistency- Based Feature Selection (CON). In Attribute Selection method, used CFS Evaluator as an Attribute Evaluate and Best First as Search Method.

Out of 42 features only 12 features were selected that are protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment, lroot_shell, count, diff_srv_rate, dst_host_same_src_port_rate, label.

E) C4.5 Decision Tree Algorithm

It is a classification algorithm in data mining. It is an induction algorithm that creates a model using the data set and deduce out some assumption. Based on the assumptions deduced, it tries to classify the new data set. It is also called as Classification tree algorithm.

In this algorithm there is a root node, and then comes internal nodes on which the tests are performed. On getting the result we reached to the leaf node which describes the ultimate result.

On the base of attributes, given data items are classified by the decision tree algorithm. Initially a decision tree is constructed with the help of pre-classified data set. Each and every data item has some set of attributes, which has some value on which they are defined.

Selection of attribute is the key issue, as we have to select the best attribute that divides the data item into corresponding classes. Partitioning of data item is made on the basis on the values of the attributes of the data element. This process is applied on every partition of data items. When all the data items are categorized together that is of same class the process gets terminated. At the end the name of leaf node is the result of classification.

C4.5 algorithm can deal with continuous attributes, missing attributes value, and gives computational efficiency.

Nodes, leaves and edges make a decision tree. Node describes that attributes on the basis on which the partitioning of the data takes place.

Every node comprises of several edges. According to the values of edges, values of attributes in parent node, the labeling is done. Two nodes or node and leaf are joined together with an edge.

Fig. 3 shows the basic and common example of decision of playing depending upon the conditions of weather.

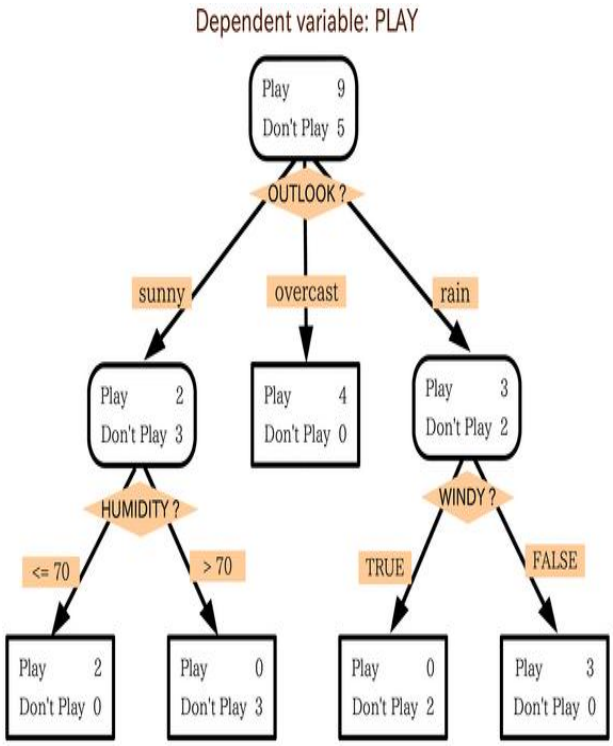


Fig- 3 Decision tree of weather data

Table 4 showing the list of parameters used in algorithm with their description and values.

Table 4- C4.5 Parameters

Parameters	Description	Values
binarySplits	On nominal attributes, whether to use binary splits.	True
unpruned	Pruning to be done.	False
collapseTree	Remove that parts that do not reduce training error.	True
subtreeRaising	When pruning whether to consider the subtree raising operation.	True
confidenceFactor	Required for pruning (smaller values incur more pruning).	0.25
seed	reduced-error pruning is used, the seed used for randomizing the data.	1
debug	Additional info to the console if true.	False
numFolds	For reduced-error pruning the amount of data used	3
minNumObj	The minimum number of instances per leaf.	5
reducedErrorPruning	Reduced error pruning or C4.5 pruning is used.	False

F) Support Vector Machine

It is a binary classification that is used for the categorizing of the attacks. If we merge the binary classifier with the decision tree algorithm then we have multi class SVM. With the help of multi class SVM we can classify attacks of different class. SVM uses non-linear mapping that maps the real values into higher dimensional feature space. Linear separating hyper plane is used by SVM for the creation of classifier. Through the use of hyper-plane SVM separate the data into different classes.There is an attribute that is called as kernel that SVM uses for solving the problem. User has to provide the kernel function at the training phase of the algorithm. With the help of support vectors, SVM does the classification. There are many kernel functions like linear, radial basis functions, polynomial, sigmoid.

In the Fig. 4 given below, the distance between the data and hyper plane is revealed. In the left image, the distance among the data and the hyper plane is small and in the right image the space is larger, which makes classification easy.

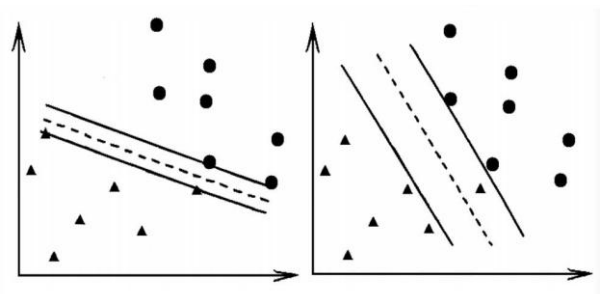


Fig 4- Hyperplane of Support Vector Machine Separating the two different Classes

Parameters that were involved in the SVM algorithm are shown below in Table 5 with their corresponding values.

Table 5- SVM Parameters

Parameters	Description	Values
SVM Type	Type of SVM used	C-SVM
shrinking	Shrinking heuristic is used	True
cacheSize	Cache size in MB	100
probabilityEstimates	Whether to generate probability estimates instead of -1/+1 for classification problem	False
Coef0	Coefficient used	0.0
nu	The value of nu for nu-SVC, one-class SVM, nu-SVR	0.5
Cost	C-SVC cost parameter	1.0
Normalise	Normalise the data	False
Debug	Additional info to the console if true.	False
kernelType	Kernel type used	Linear
Degree	The degree of the kernel	3
gamma	1/ max_index is used if 0.	0.0
eps	The toleration of the termination criterion	0.0010
DonotReplaceMissingValues	missing values automatic replacement	False

Type of SVM used is C-SVM. The data set is having two classes i.e. Abnormal and Normal.

Cache size is set to 100.0 Mb. It determines the memory in Mb given to the algorithm in RAM.

Additional information will be displayed on the console window if the Debug parameter is set to true.

Normalize option is set to false as the data set is already normalized. Probability Estimate factor is also set to false.

These are parameters that are important, repeatedly performing the experiment will bring out the best set of options.

IV. Result and Discussions

The outcomes of the implementation of the algorithm are shown below. First, the algorithms are trained with the preprocessed dataset. Dataset was separated into two parts. Through the first part, the model was prepared and with the remaining of the dataset, the model was tested. Following is the output of the process:-

Time and number of leaves produced by C4.5 algorithm are shown in Table 6.

Table 6 Results of C4.5 Algorithm

Parameters	Value
Number of Leaves	77
Size of the tree	153
Time taken to build model	49.67seconds

Accuracy of the C4.5 algorithm in percentage is shown in the Table 7.

Table 7 Accuracy of the C4.5 Algorithm

Parameters	Value	Percentage
Correctly Classified Instances	246896	99.9538 %
Incorrectly Classified Instances	114	0.0462 %
Coverage of cases (0.95 level)	99.9688 %	
Total Number of Instances	247010	

Table 8 showing the detailed accuracy by class i.e. Anomaly and Normal and the corresponding confusion Matrix obtained is show below in Table 9.

Table 8 Accuracy by Class of C4.5

TP Rate	FP Rate	Class
1	0.001	Anomaly
0.999	0	Normal

Table 9 Confusion Matrix of C4.5

a	b	<-- classified as
198223	83	| a = normal
31	48673	| b = anomaly

Now the information is added to the dataset and again it is passed to the SVM algorithm and below is the final output:-

Accuracy of the SVM algorithm in percentage is shown in the Table 10.

Table 10 Accuracy of SVM

Parameters	Value	Percentage
Correctly Classified Instances	246539	99.8093 %
Incorrectly Classified Instances	471	0.1907 %
Coverage of cases (0.95 level)	99.8093 %	
Total Number of Instances	247010	

Table 11 showing the detailed accuracy by class i.e. Anomaly and Normal and corresponding confusion Matrix obtained is show below in Table 12.

Table 11 Accuracy by Class of SVM

TP Rate	FP Rate	Class
0.998	0.001	Anomaly
0.999	0.002	Normal

Table 12 Confusion Matrix of SVM

a	b	<-- classified as
197860	446	| a = normal
25	48679	| b = anomaly

V. Conclusion and Future Scope of Work

An intrusion detection system structure is proposed, and for estimating the performance, network data was required. Since the data collection for training and evaluating the classifier is a nontrivial task and our major was to promise the uprightness of the computer system, hence NSL KDD intrusion dataset is used for estimating the system. This system framework combines two classification algorithms as a core technique. After the testing is performed on NSL KDD dataset, numerical results demonstrate that our system have slight advantage over the KDD Cup 99. False alarm rate is low and high accuracy and less time is required by the proposed architecture. However attacks were not labeled, so our system only categorizes the connection as an abnormal or normal.

Furthermore, in future some more exploration can be done on this area. Some other feature selection algorithm can be used that can select the more significant feature and make system more effective. Dataset should be collected for testing the system. System should be trained with the new dataset regularly so that it becomes capable of recognizing the new attacks. Algorithms can be tested with different set of options in order to achieve more effective results.

References

[1] R. Heady, G. Luger, A. Maccabe, M. Servilla, "The architecture of a network level intrusion detection system", Technical report, Computer Science Department, University of New Mexico, August 1990

[2] M. Mahoney, Computer security: A survey of attacks and defences, 2000, http://www.cs.fit.edu/~mmahoney/ids.html (Accessed on 9th February 2012).

[3] S. L. Scott, "A Bayesian paradigm for designing Intrusion Detection Systems", Computational Statistics & Data Analysis, 2004, 45: p. 69–83.

[4] G. Giacinto, F. Roli, L. Didaci, "Fusion of multiple classifiers for intrusion detection in computer networks", Pattern Recognition Letters, 2003, 24: p. 1795–1803.

[5] G. Kou, Y. Peng, Z. Chen, Y. Shi, "Multiple criteria mathematical programming for multi-class classification and application in network intrusion detection", Information Sciences, 2009, 179: p. 371–381.

[6] I. Kang, M. K. Jeong, D. Kong, "A differentiated one-class classification method with applications to intrusion detection", Expert Systems with Applications, 2012, 39: p. 3899-3905.

[7] S. Jiang, X. Song, H. Wang, J. Han, Q. Li," A clustering-based method for unsupervised intrusion detections", Pattern Recognition Letters, 2006, 27: p 802–810.

[8] S. Lee, G. Kim, S. Kim, "Self-adaptive and dynamic clustering for online anomaly detection", Expert Systems with Applications, 2011, 38: p. 14891–14898.

[9] V. Nikulin, "Threshold-based clustering with merging and regularization in application to network intrusion detection", Computational Statistics & Data Analysis, 2006, 51: p. 1184 – 1196.
[10] A. Tajbakhsh, M. Rahmati, A. Mirzaei, "Intrusion detection using fuzzy association rules", Applied Soft Computing, 2009, 9: p. 462–469.
[11] J. E. Dickerson and J. A. Dickerson, "Fuzzy Network Profiling for Intrusion Detection", Proceedings of NAFIPS 19th International Conference of the North American Fuzzy Information Processing Society, Atlanta, 2000, 3: p 301-306.
[12] Y. Liu, K. Chen, X. Liao, W. Zhang," A genetic clustering method for intrusion detection", Pattern Recognition, 2004, 5: p. 927–942.
[13] A. N. Toosi, M. Kahani, "A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers", Computer Communications, 2007, 30: p. 2201–2212.
[14] K. Shafi, H. A. Abbass, "An adaptive genetic-based signature learning system for intrusion detection", Expert Systems with Applications, 2009, 36: p. 12036–12043.
[15] M. S. Abadeh, H. Mohamadi, J. Habibi, "Design and analysis of genetic fuzzy systems for intrusion detection in computer networks", Expert Systems with Applications, 2011, 38: p. 7067–7075.
[16] C. Tsang, S. Kwong, H. Wang, "Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection", Pattern Recognition, 2007 40: p. 2373 – 2391.
[17] P. Sangkatsanee, N. Wattanapongsakorn, C. Charnsripinyo," Practical real-time intrusion detection using machine learning approaches", Computer Communications, 2011, 34: p. 2227-2235.
[18] Y. Yi, J. Wu, W. Xu, "Incremental SVM based on reserved set for network intrusion detection", Expert Systems with Applications, 2011, 38: p. 7698–7707.
[19] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, S. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusion in unlabelled data", Data Mining for Security Applications, Kluwer, 2002.
[20] A. K. Ghosh, A. Schwartzbard, M. Schatz," Learning program behaviour profiles for intrusion detection", Proceedings of the Workshop on Intrusion Detection and Network Monitoring, Santa Clara, California, USA, 1999, p: 9-12.
[21] C. Zhang, J. Jiang, M. Kamel, "Intrusion detection using hierarchical neural networks", Pattern Recognition Letters, 2005, 26: p. 779–791.
[22] J. Li, G. Zhang, G. Gu, "The research and implement of intelligent intrusion detection system based on artificial neural network" , Proceedings of the Third International Conference on Machine Laming and Cybernetics, Shanghai, 2004, p. 26-29.
[23] Mohammed Theeb Alotaibi, "Intelligent U2R Attack Detection Using Neural Network", M.Tech Dessertation King Saud University College of Computer and Information Sciences, 2006.
[24] D. Farid, N. Harbi, E. Bahri, M. Z. Rahman, C. M. Rahman, "Attacks Classification in Adaptive Intrusion Detection using Decision Tree", Proceeding of the International Conference on Computer Science (ICCS), Rio De Janeiro, Brazil, 2010, 63: p. 86-90.
[25] W. Lee, S. J. Stolfo et al, "A data mining and CIDF based approach for detecting novel and distributed intrusions", Lecture Notes in Computer Science, 2000, 1907: p. 49-65.
[26] Li Hanguang, Ni Yu, "Intrusion Detection Technology Research Based on Apriori Algorithm", 2012, 24: p., 1615-1620.
[27] G. Wang, J. Hao, J. Ma, L. Huang," A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering", Expert Systems with Applications, 2010, 37: p. 6225–6232.
[28] S. Horng, M. Su, Y. Chen, T. Kao, R. Chen, J. Lai, C. Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications, 2011, 38: p. 306–313.
[29] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method", Expert Systems with Applications, 2011, 39: p. 424-430.
[30] O.Depren, M. Topallar, E. Anarim, M. K. Ciliz," An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks", Expert Systems with Applications, 2005, 29: p. 713–722.
[31] T. Pietraszek, A. Tanner, "Data mining and machine learning dTowards reducing false positives in intrusion detection", Information Security Technical Report, 2005, 10: p. 169-183.
[32] S. Wu, E.Yen, "Data mining-based intrusion detectors", Expert Systems with Applications, 2009, 36: p. 5605-5612.
[33] T. Ozyer, R. Alhajj, Ken Barker," Intrusion detection by integrating boosting genetic fuzzy classifier and data mining criteria for rule pre-screening", Journal of Network and Computer Applications, 2007, 30: p. 99 – 113.
[34] S. Peddabachigaria, A. Abrahamb, C. G. J. Thomas, "Modeling intrusion detection system using hybrid intelligent systems", Journal of Network and Computer Applications, 2007, 30, 114-132.
[35] M. Ali Aydın, A. H. Zaim, K. G. Ceylan, "A hybrid intrusion detection system design for computer network security", Computers and Electrical Engineering, 2009, 35: p. 517–526.
[36] S. S. S. Sindhu, S. Geetha, A. Kannan, "Decision tree based light weight intrusion detection using a wrapper approach", Expert Systems with Applications, 2012, 39: p. 129–141.
[37] S. lakhina, S.Joseph, B. Verma, "Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD",

International Journal of Engineering Science and Technology, 2010, 2(6): p. 1790-1799.
[38] Z. Pan, S. Chen, G. Hu, D. Zhang, "Hybrid neural network and C4.5 for misuse detection", The 2nd International Conference on Machine Learning and Cybernetics, Xi'an, 2003, 4: p. 2463–2467.
[39] A. Chittur,"Model generation for an intrusion detection system using genetic algorithms", High School Honors Thesis, Ossining High School, in cooperation with Columbia Univ, 2001.
[40] M. Su, "Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbour classifiers", Expert Systems with Applications, 2011, 38: p. 3492–3498.
[41] M. Jiang, "Combining Multiple Techniques for Intrusion Detection", International Journal of Computer Science and Network Security, 2006, 6: p. 208-218.
[42] M. Panda, A. Abraham, M.R. Patra, "Discriminative multinomial Naïve Bayes for network intrusion detection", Proceedings of the 6th International Conference on Information Assurance and Security (IAS), 2010, p. 5-10.
[43] A. P. Muniyandi, R. Rajeswari, R. Rajaram, "Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm", Procedia Engineering, 2012, 30:p. 174-182.
[44] I.H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, S. J. Cunningham, "Weka: Practical Machine Learning Tools and Techniques with Java Implementations", Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems, Dunedin, New Zealand, 1999, p. 192-196.
[45] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", Proceedings of the 2009 IEEE Symposium on Computational Intelligence, Ottawa, Canada, 2009, p. 53-58.

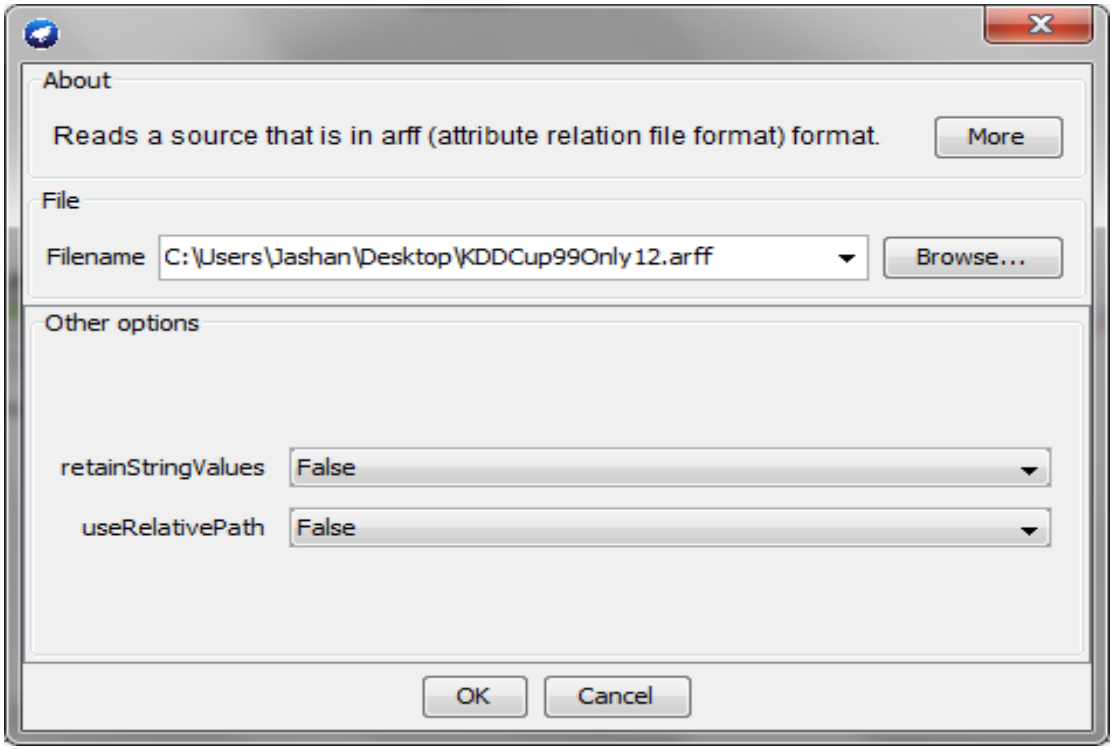
Jashan Koshal received the Bachelor of Engineering degree in Information Technology from Jabalpur Engineering College, Jabalpur, India in 2010. He is currently pursuing his Master of Technology in Information Technology with specialization in Software Engineering from Indian Institute of Information Technology, Allahabad, India. His general research interest is in the area of network security, intrusion detection and data mining.

Monark Bag is a Lecturer in MBA (IT) and MS (CLIS) Division of Indian Institute of Information Technology, Allahabad. He holds a B.Tech (Computer Science and Engineering), MBA (Information Technology Management) and PhD (Engineering). He is highly engaged in teaching and research. His research interest includes expert system, control chart pattern recognition, quality control, optimization techniques and intrusion detection systems. He has published many papers in reputed journals, conferences and book chapters.

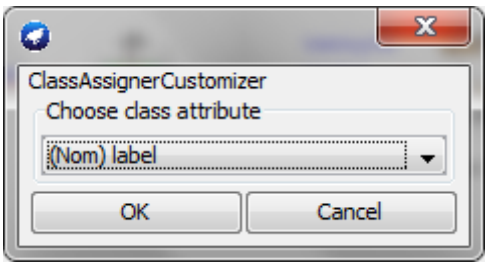
APPENDIX -1

Initial screen of the software Weka.

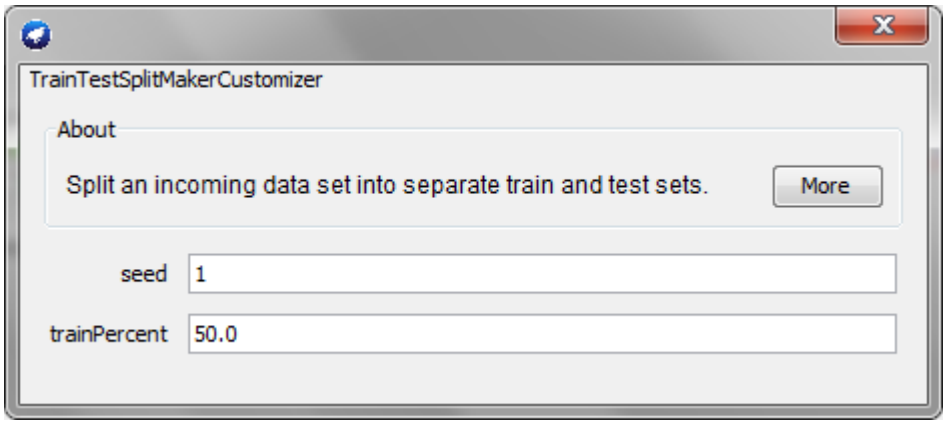
APPENDIX- 2
Loading of the data set



APPENDIX- 3
Assigning the Class Attribute



APPENDIX- 4
Splitting the dataset for training and testing respectively



APPENDIX- 5
Algorithms are applied on the dataset