A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems

Reema Patel, Amit Thakkar, Amit Ganatra

Abstract*— Despite of growing information technology widely, security has remained one challenging area for computers and networks. In information security, intrusion detection is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource. Currently many researchers have focused on intrusion detection system based on data mining techniques as an efficient artifice. Data mining is one of the technologies applied to intrusion detection to invent a new pattern from the massive network data as well as to reduce the strain of the manual compilations of the intrusion and normal behavior patterns. This article reviews the current state of art data mining techniques, compares various data mining techniques used to implement an intrusion detection system such as Decision Trees, Artificial Neural Network, Naïve Bayes, Support Vector Machine and K- Nearest Neighbour Algorithm by highlighting advantages and disadvantages of each of the techniques. Finally, a discussion of the future technologies and methodologies which promise to enhance the ability of computer systems to detect intrusion is provided and current research challenges are pointed out in the field of intrusion detection system.

***Index Terms*— Classification, Data Mining, Intrusion Detection System**

I. INTRODUCTION

Intrusion detection technique is technology designed to observe computer activities for the purpose of finding security violations. The security of a computer system is compromised when an intrusion takes place. Intrusion detection is the process of identifying and responding to malicious activity targeted at computing and networking sources [1]. Intrusion prevention techniques, such as user authentication and information protection have been used to protect computer systems as a first line of defense. Intrusion prevention alone is not sufficient because as systems become ever more complex, there are always exploitable weaknesses in the systems due to design and programming errors. Now a day, intrusion detection is one of the high priority tasks for network administrators and security professionals.

As network based computer systems play increasingly vital roles in modern society, they have become intrusion detection systems provide following three essential security functions:

- ***Data confidentiality:*** Information that is being transferred through the network should be accessible only to those that have been properly authorized.
- ***Data integrity:*** Information should maintain their integrity from the moment they are transmitted to the moment they are actually received. No corruption or data loss is accepted either from the random events or malicious activity.
- ***Data availability:*** The network or a system resource that ensures that it is accessible and usable upon demand by an authorized system user.

Any intrusion detection system has some inherent requirements. Its prime purpose is to detect as many attacks as possible with minimum number of false alarms, i.e. the system must be accurate in detecting attacks. However, an accurate system that cannot handle large amount of network traffic and is slow in decision making will not fulfill the purpose of an intrusion detection system (IDs). Data mining techniques like data reduction, data classification, features selection techniques play an important role in IDS.

This work is a survey of data mining techniques that have been applied to IDSs and is organized as follows: Section 2 presents IDs terminology and taxonomy. Section 3 mentions the drawbacks of standard IDs. Section 4 gives brief introduction about data mining .Section 5 illustrates how data mining can be used to enhance IDSs. Section 6 describes the various data mining approaches that have been employed in IDSs by various researchers. Section 7 provides misuse and anomaly detection using data mining techniques. Section 8 describes various data mining algorithms to implement IDs and also compares various data mining algorithms that are being used to implement IDs. Section 9 provides experimental study on weka environment. Section 10 focuses on current research challenges and finally section 10 concludes the work.

II. IDs TERMINLOGY AND TAXONOMY

IDs uses several techniques to determine what qualifies as an intrusion versus normal traffic. There are two useful method of classification for intrusion detection systems is according to data source. Each has a distinct approach for monitoring, securing data and systems. There are two following general categories under this classification:

Manuscript received February 2012

Reema Patel, Department of Information Technology, Charotar University of Science and Technology, Changa 388421, Anand, Gujarat, (e-mail: reemapatel.it@ ecchanga.ac.in).

Amit Thakkar, Department of Information Technology, Charotar University of Science and Technology, Changa 388421, Anand, Gujarat, (e-mail: amitthakkar.it@ecchanga.ac.in).

Amit Ganatra, U and P U Patel Department of Computer Engineering, Charotar University of Science and Technology, Changa 388421, Anand, Gujarat, (e-mail: amitganatra.ce@ecchanga.ac.in)

- **Host-based IDSs (HIDS)** – examine data held on individual computers that serve as hosts. The network architecture of host-based is agent-based, which means that a software agent resides on each of the hosts that will be governed by the system [12].
- **Network-based IDSs (NIDS)** – examine data exchanged between computers. Most efficient host-based intrusion detection systems are capable of monitoring and collecting system audit in real time as well as on a scheduled basis, thus distributing both CPU utilization and network overhead and providing for a flexible means of security administration [12].

A. Intrusion Detection Approaches

The signatures of some attacks are known, whereas other attacks only reflect some deviation from normal patterns. Consequently, two main approaches have been devised to detect intruders.

Anomaly Detection: Anomaly detection assumes that intrusions will always reflect some deviations from normal patterns. Anomaly detection may be divided into static and dynamic anomaly detection. A static anomaly detector is based on the assumption that there is a portion of the system being monitored that does not change. The static portion of a system is the code for the system and the constant portion of data upon which the correct functioning of the system depends. For example, the operating systems' software and data to bootstrap a computer never change. If the static portion of the system ever deviates from its original form, an error has occurred or an intruder has altered the static portion of the system. Dynamic anomaly detection typically operates on audit records or on monitored networked traffic data. Audit records of operating systems do not record all events; they only record events of interest. Therefore only behaviour that results in an event that is recorded in the audit will be observed and these events may occur in a sequence.

Misuse Detection: It is based on the knowledge of system vulnerabilities and known attack patterns. Misuse detection is concerned with finding intruders who are attempting to break into a system by exploiting some known vulnerability. Ideally, a system security administrator should be aware of all the known vulnerabilities and eliminate them. The term intrusion scenario is used as a description of a known kind of intrusion; it is a sequence of events that would result in an intrusion without some outside preventive intervention. An intrusion detection system continually compares recent activity to known intrusion scenarios to ensure that one or more attackers are not attempting to exploit known vulnerabilities. To perform this, each intrusion scenario must be described or modeled.

B. Advantages and Disadvantages of Anomaly Detection and Misuse Detection

The main disadvantage of misuse detection approaches is that they will detect only the attacks for which they are trained to detect. Novel attacks or unknown attacks or even variants of common attacks often go undetected. The main advantage of anomaly detection approaches is the ability to detect novel attacks or unknown attacks against software systems, variants of known attacks, and deviations of normal usage of programs regardless of whether the source is a privileged internal user or an unauthorized external user. The disadvantage of the anomaly detection approach is that well-known attacks may not be detected, particularly if they fit the established profile of the user. Once detected, it is often difficult to characterize the nature of the attack for forensic purposes. Finally a high false positive rate may result for a narrowly trained detection algorithm, or conversely, a high false negative rate may result for a broadly trained anomaly detection approach.

C. Combining misuse and anomaly detection

Anomaly detection and misuse detection have major shortcomings that hamper their effectiveness in detecting intrusions. Research can be carried into intrusion detection methodologies which combine the anomaly detection approach and the misuse detection approach [2]. These techniques seek to incorporate the benefits of both of the standard approaches to intrusion detection. The combined approach permits a single intrusion detection system to monitor for indications of external and internal attacks.

While a significant advantage over the singular use of either method separately, the use of a combined anomaly/misuse mechanism does possess some disadvantages. The use of two knowledgebase for the intrusion detection system will increase the amount of system resources which must be dedicated to the system [3]. Additional disk space will be required for the storage of the profiles, and increased memory requirements will be encountered as the mechanism compares user activities with information in the dual knowledge bases. In addition, the technique will share the disadvantage of either method individually in its inability to detect collaborative or extended attack scenarios.

Pattern recognition possesses a distinct advantage over anomaly and misuse detection methods in that it is capable of identifying attacks which may occur over an extended period of time, a series of user sessions, or by multiple attackers working in concert. This approach is effective in reducing the need to review a potentially large amount of audit data [3].

III. DRAWBACKS OF IDSs

Intrusion Detection Systems (IDS) have become a standard component in security infrastructures as they allow networks administrators to detect policy variations. These policy violations range from external attackers trying to gain unauthorized access to intruders abusing their access. Current IDS have a number of significant drawbacks [4]:

- **False positives:** A common complaint is the amount of false positives an IDS will generate. Developing unique signatures is a difficult task. It is much more difficult to pick out a valid intrusion attempt if a signature also alerts regularly on valid network activity.
- **False negatives:** Detecting attacks for which there are no known signatures. This leads to the other concept of false negatives where AN ID does not generate an alert when an intrusion is actually taking place. Simply put if a signature has not been written for a particular exploit there is an extremely good chance that the IDS will not detect it**.**
- **Data overload:** Another aspect which does not relate directly to misuse detection but it is extremely important is how much data an analyst can effectively and efficiently analyze.

Data mining can help to improve intrusion detection by addressing each and every one of the above mentioned problems. To accomplish these tasks, data miners employ one or more of the following techniques:

- Data summarization with statistics, including finding outliers
- Visualization: presenting a graphical summary of the data
- Clustering of the data into natural categories
- Association rule discovery: defining normal activity and enabling the discovery of anomalies
- Classification: predicting the category to which a particular record belongs.

IV. DATA MINING - INTRODUCTION

Data mining refers to a process of non-trivial extraction of implicit, previously unknown, and potentially useful information from data. It is a convenient way of extracting patterns, which represents mining implicitly stored in large data sets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability. It can be viewed as an essential step in the process of knowledge data discovery.

Here are a few specific things that data mining might contribute to an intrusion detection system:

- Remove normal activity from alarm data to allow analysis to focus on real attacks.
- Identify false alarm generators and “bad” sensor signatures
- Find anomalous activity that uncovers a real attack.
- Identify long, ongoing patterns (different IP address, same activity)

Benefits of Data Mining Techniques

1. Problems with large databases may contain valuable implicit regularities that can be discovered automatically.

2. Difficult-to-program applications, which are too difficult for traditional manual programming.

3. Software applications that customize to the individual user’s preferences, such as personalized advertising.

There are several reasons why data mining approaches plays a role in these three domains. First of all, for the classification of security incidents, a vast amount of data has to be analyzed containing historical data. It is difficult for human beings to find a pattern in such an enormous amount of data. Data mining, however, seems well-suited to overcome this problem and can therefore be used to discover those patterns.

Reasons to use data mining approaches in IDS

1. It is very hard to program an IDS using ordinary programming languages that require the explicitation and formalization of knowledge.
2. The adaptive and dynamic nature of machine-learning makes it a suitable solution for this situation.
3. The environment of an IDS and its classification task highly depend on personal preferences. What may seem to be an incident in one environment may be normal in other environments. This way, the ability of computers to learn enables them to know someone’s “personal” (or organizational) preferences, and improve the performance of the IDS, for this particular environment [7].

V. THE DATA MINING PROCESS OF BUILDING INTRUSION DETECTION MODELS

With the recent rapid development in KDD, a better understanding of the techniques and process frameworks that can support systematic data analysis on the vast amount of audit data that can be made available. The process of using data mining approaches to build intrusion detection models is shown in Fig 1.

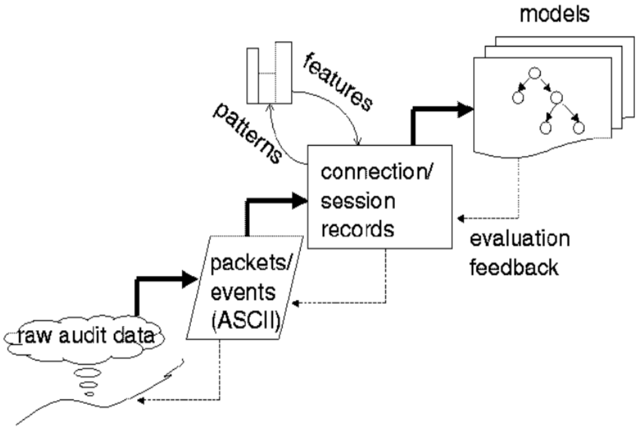

Fig (a): The Data Mining Process of Building ID Models

Here raw (binary) audit data is first processed into ASCII network packet information (or host event data), which is in turn summarized into connection records (or host session records) containing a number of within-connection features, e.g., service, duration, flag (indicating the normal or error status according to the protocols), etc. Data mining programs are then applied to the connection records to compute the frequent patterns, i.e., association rules and frequent episodes, which are then analyzed to construct additional features for the connection records. Classification programs, for example, RIPPER, are then used to inductively learn the detection models. This process is of course iterative. For example, poor performance of the classification models often indicates that more pattern mining and feature construction is needed [5].

VI. DATA MINING APPROACHES FOR IDS

The central theme of our approach is to apply data mining techniques for intrusion detection in email system Internet-based. Data mining generally refers to the process of (automatically) extracting models from large stores of data [8]. The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning, and database. Several types of algorithms [8] are particularly relevant to our research:

Classification: Maps a data item into one of several pre-defined categories. These algorithms normally out-put “classifiers”, for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient “normal” and “abnormal” audit data for a user or a program, then apply a classification algorithm to learn a classifier that can label or predict new unseen audit data as belonging to the normal class or the abnormal class [6].

Link analysis: Determines relations between fields in the database. Finding out the correlations in audit data will provide insight for selecting the right set of system features for intrusion detection [6].

Sequence analysis: Models sequential patterns. These algorithms can discover what time-based sequence of audit events are frequently occurring together [6]. These frequent event patterns provide guidelines for incorporating temporal statistical measures into intrusion detection models. For example, patterns from audit data containing network-based denial-of-service (DOS) attacks suggest that several per-host and per-service measures should be included.

VII. MISUSE AND ANOMALY DETECTION USING DATA MINING TECHNIQUES

A. Misuse Detection Using Supervised Learning:

Misuse detection methods, a model based supervised method make use of a classifier that has to be trained with labeled patterns [7]. The training patterns are labeled as 'normal' or 'attacks'. After the classifier is trained, it can classify or label new unlabeled patterns. These methods are also able to detect previously known attacks with good accuracy but also have some disadvantages. They are unable to detect new emerging threats and the labeling procedure of the training data is expensive and time consuming.

B. Anomaly Detection Using Supervised Learning:

The supervised anomaly detection approach train a classifier with pure "normal" labeled patterns. Anomalies (a subset of which is attacks) are detected as significant deviations from this model of normal behavior. The arguments for this approach are that normal data is far easier to come by than are labeled attacks that a pure anomaly detector is unbiased towards any set of pre-trained attacks, and, therefore, it may be capable of detecting completely novel attacks. The counter arguments are that hostile activities which appear similar to normal behavior are likely to go undetected, that it fails to exploit prior knowledge about a great many known attacks, and that, to date, false alarm rates for pure anomaly detection systems remain unusable high.

C. Misuse Detection Using Unsupervised Learning:

As is known unsupervised learning is based not on the predefined training data set misuse detection is done mostly by using supervised learning and the unsupervised learning is not been preferred for misuse detection .

D. Anomaly Detection Using Unsupervised Learning:

The unsupervised anomaly detection approach overcome the problem of labeling procedure of the training data that is very expensive and time consuming by making use of data clustering algorithms, which makes no assumption about the labels or classes of the patterns. The patterns are grouped together based on a similarity measure and the anomalies or attacks are the patterns in the smaller clusters. Two assumptions need to be made for this to be true: the normal patterns or connections are many more than the attacks and that the attacks are different than the normal patterns. The drawback of data clustering for anomaly detection is a potential false alarm rate [7].

VIII. DATA MINING ALGORITHMS TO IMPLEMENT INTRUSION DETCTION SYSTEM

Data mining algorithms automatically extract knowledge from machine readable information. In data mining, computer algorithms (learners) attempt to automatically distill knowledge from example data. This knowledge can be used to make predictions about novel data in the future and to provide insight into the nature of the target concepts.

Most Popular Data Mining Algorithms for IDs

Bayes Classifier:

A Bayesian network is a model that encodes probabilistic relationships among variables of interest. This technique is generally used for intrusion detection in combination with statistical schemes, a procedure that yields several advantages, including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data. However, a serious disadvantage of using Bayesian networks is that their results are similar to those derived from threshold-based systems, while considerably higher computational effort is required.

K-Nearest Neighbour:

K-Nearest Neighbour (k-NN) is instance based learning for classifying objects based on closest training examples in the feature space. It is a type of lazy learning where the function is only approximated locally and all computation s deferred until classification. The k-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbors. If k=1, then the object is simply assigned to the class of its nearest neighbor. The k-NN algorithm uses all labeled training instances as a model of the target function. During the classification phase, k-NN uses a similarity-based search strategy to determine a locally optimal hypothesis function. Test instances are compared to the stored instances and are assigned the same class label as the k most similar stored instances. Generally it is used for intrusion detection in combination with statistical schemes (anomaly detection).

Decision Tree:

Decision tree is a predictive modeling technique most often used for classification in data mining. The Classification algorithm is inductively learned to construct a model from the preclassified data set. Each data item is defined by values of the attributes. Classification may be viewed as mapping from a set of attributes to a particular class. The Decision tree classifies the given data item using the values of its attributes. The decision tree is initially constructed from a set of pre-classified data. The main approach is to select the attributes, which best divides the data items into their classes. According to the values of these attributes the data items are partitioned. This process is recursively applied to each partitioned subset of the data items. The process terminates when all the data items in current subset belongs to the same class. A node of a decision tree specifies an attribute by which the data is to be partitioned. Each node has a number of edges, which are labeled according to a possible value of the attribute in the parent node. An edge connects either two nodes or a node and a leaf. Leaves are labeled with a decision value for categorization of the data.

Induction of the decision tree uses the training data, which is described in terms of the attributes. The main problem here is deciding the attribute, which will best partition the data into various classes. To classify an unknown object, one starts at the root of the decision tree and follows the branch indicated by the outcome of each test until a leaf node is reached. The name of the class at the leaf node is the resulting classification.

Decision trees can be used as a misuse intrusion detection as they can learn a model based on the training data and can predict the future data as one of the attack types or normal based on the learned model. Decision trees work well with large data sets. This is important as large amounts of data flow across computer networks. The high performance of Decision trees makes them useful in real-time intrusion detection. Decision trees construct easily interpretable models, which is useful for a security officer to inspect and edit. These models can also be used in the rule-based models with minimum processing. Generalization accuracy of decision trees is another useful property for intrusion detection model. There will always be some new attacks on the system which are small variations of known attacks after the intrusion detection models are built. The ability to detect these new intrusions is possible due to the generalization accuracy of decision trees.

Neural Network (NN):

Neural networks have been used both in anomaly intrusion detection as well as in misuse intrusion detection. For anomaly intrusion detection, neural networks were modeled to learn the typical characteristics of system users and identify statistically significant variations from the user's established behavior. In misuse intrusion detection the neural network would receive data from the network stream and analyze the information for instances of misuse. A NN for misuse detection is implemented [9] in two ways. The first approach incorporates the neural network component into an existing or modified expert system. This method uses the neural network to filter the incoming data for suspicious events and forward them to the expert system. This improves the effectiveness of the detection system. The second approach uses the neural network as a stand alone misuse detection system. In this method, the neural network would receive data from the network stream and analyze it for misuse intrusion. There are several advantages to this approach. It has the ability to learn the characteristics of misuse attacks and identify instances that are unlike any which have been observed before by the network. It has high degree of accuracy to recognize known suspicious events. Generally, it is used to learn complex non linear input-output relationships.

Support Vector Machine:

Support Vector Machines [10] have been proposed as a novel technique for intrusion detection.An SVM maps input (real-valued) feature vectors into a higher-dimensiona feature space through some nonlinear mapping. SVMs are developed on the principle of structural risk minimization [11]. Structural risk minimization seeks to find a hypothesis h for which one can find lowest probability of error whereas the traditional learning techniques for pattern recognition are based on the minimization of the empirical risk, which attempt to optimize the performance of the learning set. Computing the hyper plane to separate the data points i.e. training an SVM leads to a quadratic optimization problem. SVM uses a linear separating hyper plane to create a classifier but all the problems cannot be separated linearly in the original input space. SVM uses a feature called kernel to solve this problem. The Kernel transforms linear algorithms into nonlinear ones via a map into feature spaces. There are many kernel functions; including polynomial, radial basis functions, two layer sigmoid neural nets etc. The user may provide one of these functions at the time of training the classifier, which selects support vectors along the surface of this function. SVMs classify data by using these support vectors, which are members of the set of training inputs that outline a hyper plane in feature space.

Computing the hyper plane to separate the data points i.e. training a SVM leads to quadratic optimization problem. SVM uses a feature called kernel to solve this problem. Kernel transforms linear algorithms into nonlinear ones via a map into feature spaces. There are many kernel functions; some of them are Polynomial, radial basis functions, two layer sigmoid neural nets etc. The user may provide one of these functions at the time of training classifier, which selects support vectors along the surface of this function. SVMs classify data by using these support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. The implementation of SVM intrusion detection system has two phases: training and testing.

SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification. The main disadvantage is SVM can only handle binary-class classification whereas intrusion detection requires multi-class classification

IX. EXPERIMENTAL METHODOLOGY

A. The Data Set

The data set used for the entire course of research is the DARPA KDD99 benchmark data set [13], also known as "DARPA Intrusion Detection Evaluation data set" that not only includes a large quantity of network traffic but also collects a wide variety of attacks. Attack fall into following four main classes:

- **Denial of service (DoS) attacks:** Attackers disrupt a host or network service to make legitimate users can not access to a machine, e.g. ping-of-death and SYN flood;
- **Remote to Local (R2L) attacks:** Unauthorized attackers gain local access from a remote machine and then exploit the machine's vulnerabilities, e.g. guessing password;
- **User to Root (U2R) attacks:** Local users get access to local machine without authorization and then exploit the machine's vulnerabilities, e.g. various "buffer overflow" attacks; and
- **Probes:** It is a category of attacks where an attacker examines a network to discover well-known vulnerabilities. These network investigations are reasonably valuable for an attacker who is staging an attack in future.

Table I: GENERAL COMPARISON CLASSIFIERS

Classifier	Method	Parameters	Advantages	Disadvantages
Support Vector Machine	A support vector machine constructs a hyper plane or set of hyper planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks.	The effectiveness of SVM lies in the selection of kernel and soft margin parameters. For kernels, different pairs of (C, γ) values are tried and the one with the best cross-validation accuracy is picked. Trying exponentially growing sequences of C is a practical method to identify good parameters.	1. Highly Accurate 2. Able to model complex nonlinear decision boundaries 3. Less prone to over fitting than other methods	1. High algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale tasks. 2. The choice of the kernel is difficult 3. The speed both in training and testing is slow.
K Nearest Neighbour	An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its *k* nearest neighbours (*k* is a positive integer). If *k* = 1, then the object is simply assigned to the class of its nearest neighbour.	Two parameters are considered to optimize the performance of the kNN, the number k of nearest neighbour and the feature space transformation.	1. Analytically tractable. 2. Simple in implementation 3. Uses local information, which can yield highly adaptive behaviour 4. Lends itself very easily to parallel implementations	1. Large storage requirements. 2. Highly susceptible to the curse of dimensionality. 3. Slow in classifying test tuples.
Artificial Neural Network	An ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.	ANN uses the cost function *C* is an important concept in learning, as it is a measure of how far away a particular solution is from an optimal solution to the problem to be solved.	1. Requires less formal statistical training. 2. Able to implicitly detect complex nonlinear relationships between dependent and independent variables. 3. High tolerance to noisy data. 4. Availability of multiple training algorithms.	1. "Black box" nature. 2. Greater computational burden. 3. Proneness to over fitting. 4. Requires long training time.
Bayesian Method	Based on the rule, using the joint probabilities of sample observations and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation.	In Bayes, all model parameters (*i.e.,* class priors and feature probability distributions) can be approximated with relative frequencies from the training set.	1. Naïve Bayesian classifier simplifies the computations. 2. Exhibit high accuracy and speed when applied to large databases.	1 The assumptions made in class conditional independence. 2. Lack of available probability data.
Decision Tree	Decision tree builds a binary classification tree. Each node corresponds to a binary predicate on one attribute; one branch corresponds to the positive instances of the predicate and the other to the negative instances.	Decision Tree Induction uses parameters like a set of candidate attributes and an attribute selection method.	1. Construction does not require any domain knowledge. 2. Can handle high dimensional data. 3. Representation is easy to understand. 4. Able to process both numerical and categorical data.	1. Output attribute must be categorical. 2. Limited to one output attribute. 3. Decision tree algorithms are unstable. 4. Trees created from numeric datasets can be complex.

Table II: COMPARISON OF DIFFERENT CLASSIFIERS ON WEKA

	Method Name	Correctly Classified Instances in Full Dataset (%)	Incorrectly Classified Instances in Full Dataset (%)
SVM	functions.SMO	63.9747	36.0253
ANN	functions.MultilayerPerceptron	65.83	34.17
KNN	lazy.IBk	51.6878	48.3122
NB	bayes.NaiveBayes	55.7722	44.2278
DT	trees.J48	52.7004	47.2996

As shown in the Table - I, there are various approaches to implement an intrusion detection system based on its type and mode of deployment. Each of the approaches to implement an intrusion detection system has its own advantages and disadvantages. This is apparent from the discussion of comparison among the various methods. Thus it is difficult to choose a particular method to implement an intrusion detection system over the other. In Table –II, results are generated using dataset in weka environment by selecting different classification algorithms. It is cleared by obtaining the results that no algorithm gives the better accuracy in attack detection.

X. ResEARCH CHALLENGES

Following are the research challenges of the existing intrusion detection classification problem using data mining technique:

1. The final decision must be wrong if the output of selected classifier is wrong.
2. The trained classifier may not be complex enough to handle the problem.

XI. Conclusion

This paper draws the conclusions on the basis of implementations performed using various data mining algorithms. Combining more than one data mining algorithms may be used to remove disadvantages of one another. Thus a combining approach has to be made while selecting a mode to implement intrusion detection system. Combining a number of trained classifiers lead to a better performance than any single classifier. Errors can be complemented by other correct classifications. Different classifiers have different knowledge regarding the problem. To decompose a complex problem into sub- problems for which the solutions obtained are simpler to understand, implement, manage and update.

References

[1] Amoroso EG (1999) Intrusion detection: an introduction to internet surveillance, correlation, trace back, traps, and response. Intrusion.Net Books, NJ

[2] Lunt, T.F. (1989). Real -Time Intrusion Detection. Proceedings from IEEE COMPCON.

[3] James Cannady, Jay Harrell (1996). A comparative Analysis of current Intrusio n Detection Technologies.

[4] (SANS: FAQ: Data Mining in Intrusion Detection) http://www.sans.org/security-resources/idfaq/data_mining.php

[5] W. Lee. A Data Mining Framework for Constructing Features and Models for Intrusion Detection Systems. PhD Thesis, Computer Science Department, Columbia University, June 1999.

[6] W. Lee and S. Stolfo. Data Mining Approaches for Intrusion Detection. In proceedings of the 7th USENIX Security Symposium, 1998.

[7]. Data Mining Machine Learning Techniques – A Study on Abnormal Anomaly Detection System. M. Sathya Narayana, B. V. V. S. Prasad,A. Srividhya,K. Pandu Ranga Reddy. Issue 6, September 2011, International Journal of Computer Science and Telecommunications, Vol. Volume 2, pp. 8-14. ISSN 2047-3338 .

[8] W. Lee, S.J. Stolfo, K.W. Mok, Algorithms for Mining System Audit Data, in Proc. KDD, 1999.

[9] J. Cannady. Artificial Neural Networks for Misuse Detection. National Information Systems Security Conference, 1998.

[10] S. Mukkamala, G. Janoski, A. Sung. Intrusion Detection Using Neural Networks and Support Vector Machines. Proceedings of IEEE International Joint Conference n Neural Networks, pp.1702-1707, 2002

[11] Valdimir V. N. The Nature of Statistical Learning Theory, Springer, 1995.

[12] G.V.Nadiammai, S.Krishaveni, M.Hemalatha – "A comprehensive Analysis and study in intrusion detection system using data mining Techniques". IJCA, Volume 35 –No.8, December 2011.

[13] KDD Cup 1999 Dataset: kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Reema Patel has received her B.E degree in Information Technology from Gujarat University, Gujarat, India in 2008. she is currently working as an Lecturer in the Department of Information Technology. She has joined M.Tech at Charotar University of Science and Technology, Changa, Gujarat, India in 2010. Her current research interest include data mining in cyber security.

Amit Thakkar has received his B.E degree in Information Technology from Gujarat University, Gujarat, India in 2002 and master Degree from Dharmsinh Desai University, Gujarat, India in 2007. He has joined his Ph.D in the area of Multi relational Classification at KadiSarvavishvidhalayaUniversity, Gandhinagar, India in June 2010. Since 2002 he has been with faculty of Engineering and Technology, Charotar University of Science and Technology, Changa, Gujarat, Where he is currently working as an Associate Professor in the Department of Information Technology. He has published more than 20 research papers in the field of data mining and web technology. His current research interest includes Multi relational Data Mining, Relational Classification and Associate Classification.

Amit P. Ganatra (B.E.-'00-M.E. '04-Ph.D.* '11) has received his B.Tech. and M.Tech. degrees in 2000 and 2004 respectively from Dept. of Computer Engineering, DDIT-Nadiad from Gujarat University and Dharmsinh Desai University, Gujarat and he is pursuing Ph.D. in Information Fusion Techniques in Data Mining from KSV University, Gandhinagar, Gujarat, India and working closely with Dr.Y.P.Kosta (Guide). He is a member of IEEE and CSI. His areas of interest include Database and Data Mining, Artificial Intelligence, System software, soft computing and software engineering. He has 11 years of teaching experience at UG level and concurrently 7 years of teaching and research experience at PG level, having good teaching and research interests. In addition he has been involved in various consultancy projects for various industries.

After spending almost a year in C.U.Shah college of Engineering, Wadhwan, Gujarat, he joined CITC as a faculty member in 2001. His general research includes Data Warehousing, Data Mining and Business Intelligence, Artificial Intelligence and Soft Computing. In these areas, he is having good research record and published and contributed over 70 papers (Author and Co-author) published in referred journals and presented in various international conferences. He has guided more than 90 industry projects at under graduate level and 47 dissertations at Post Graduate level.

He is concurrently holding Associate Professor (Jan 2010 till date), Headship in computer Engineering Department (since 2001 to till date) at CSPIT, CHARUSAT and Deanship in Faculty of Technology-CHARUSAT (since Jan 2011 to till date), Gujarat. He is a member of Board of Studies (BOS), Faculty Board and Academic Council for CHARUSAT and member of BOS for Gujarat Technological University (GTU). He was the founder head of CE and IT departments of CITC (now CSPIT).