International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 2, Issue 2, February 2013 ISSN 2319 - 4847

A Review of Data Mining based Intrusion Detection Techniques

Kamini Maheshwar[1] and Divakar Singh[2]

[1,2] Department of CSE BUIT, Barkatullah University Bhopal, (M.P), India

ABSTRACT

Traditional Data Mining techniques operate on structured data such as corporate databases; this has been an active area of research for many years. Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems. Intrusion detection is an area growing in relevance as more and more sensitive data are stored and processed in networked systems. An intrusion detection system (IDS) monitors networked devices and looks for anomalous or malicious behavior in the patterns of activity in the audit stream. A comprehensive ID requires a significant amount of human expertise and time for development. Data mining-based IDSs require less expert knowledge yet provide good performance. These systems are also capable of generalizing to new and unknown attacks. Data mining based intrusion Building an IDS is a complex task of knowledge engineering. In this paper we represent a survey of data mining based intrusion techniques. The data mining techniques are categorized based upon different approaches like association rule, classification techniques. The detection type is borrowed from intrusion detection as either misuse detection or anomaly detection. This paper provides the major advancement in the data mining based intrusion detection research using these approaches, the features and categories in the surveyed work.

Keywords- Data Mining, Association Rule, Classification Techniques, Intrusion Detection.

1. INTRODUCTION

An intrusion detection system (IDS) is a system for the detection of such intrusions. The development of IDS is motivated by the following factors:

Most existing systems have security was that render them susceptible to intrusions, and finding and fixing all these deficiencies are not feasible. Prevention techniques cannot be sufficient. It is almost impossible to have an absolutely secure system. Even the most secure systems are vulnerable to insider attacks. New intrusions continually emerge and new techniques are needed to defend against them.

Since there are always new intrusions that cannot be prevented, IDS is introduced to detect possible violations of a security policy by monitoring system activities and response. IDSs are aptly called the second line of defense, since IDS comes into the picture after an intrusion has occurred. If we detect the attack once it comes into the network, a response can be initiated to prevent or minimize the damage to the system. It also helps prevention techniques improve by providing information about intrusion techniques.

More recently with the advent of the World Wide Web, a rapidly growing repository of unstructured data (in the form of text documents) has become available. Information Retrieval is the science of searching for information in the documents. Research began in the 1980s in response to a need for automatic methods of locating documents in large collections of texts. The commercial importance of this area grew massively following the advent of the World Wide Web in 1991 and subsequent exponential growth in the number of web pages. Text Mining is the science of extracting novel, interesting and non-trivial information from text. It is a much younger field than both information retrieval and data mining, but is believed to have high commercial potential value, particularly compared to data mining due to the fact that most information is stored as text, and this area is currently largely unexploited.

Data mining techniques can be differentiated by their different model functions and representation, preference criterion, and algorithms. The main function of the model that we are interested in is classification, as normal, or malicious, or as a particular type of attack. We are also interested in link and sequence analysis. Additionally, data mining systems provide the means to easily perform data summarization and visualization, aiding the security analyst in identifying areas of concern. The models must be represented in some form. Common representations for data mining techniques include rules, decision trees, linear and non-linear functions, instance-based examples, and probability models.

In this paper we represent a survey of data mining techniques that have been applied to IDSs by various researches**.**

2. LITERATURE SURVEY

The successful data mining techniques are themselves not enough to create deployable IDSs. Despite the promise of better detection performance and generalization ability of data mining-based IDSs, there are some inherent difficulties in the implementation and deployment of these systems. In this paper, we discuss several problems inherent in developing and deploying a real-time data mining-based IDS and present an overview of our research, which addresses these problems. These problems are independent of the actual learning algorithms or models used by IDS and must be overcome in order to implement data mining methods in a deployable system [1] and [3] and [5].

2.1 Association Rule

Association rules mining identifies associations (patterns or relations) among database attributes and their values. It is a pattern-discovery technique which does not serve to solve classification problems (it does not classify samples into some target classes) nor prediction problems (it does not predict the development of the attribute values). Association rules mining generally searches for any associations among any attributes present in the database.

Association rule (AR) is commonly understood as an implication X → Y in a transaction database D = $\{t_1 \ldots\ldots. t_m\}$. Each transaction $t_i \in D$ contains a subset of items I = $\{i_1 \ldots\ldots. i_n\}$. X and Y are disjoint itemsets, it holds X; Y $\subseteq$ I and X $\cap$ Y = Φ. The left hand side of this implication is called antecedent, the right hand side is referred to as consequent. The transaction database D can also be viewed as a boolean dataset where the boolean values of attributes in records express occurrence of items in transactions.

Association rule mining problem poses the question of efficiency. The number of potential rules X → Y defined by X $\subseteq$ Ix $\in$ $\{ix_1; \ldots\ldots ; ix_n\}$, Y $\subseteq$ I_y $\in$ $\{iy_1; \ldots. ; iy_m\}$, where Ix and Iy are disjoint, is equal to $2^{(m+n)}$. When general datasets are considered, the AR mining problem is known to be NP-complete. In restricted cases, for example in sparse boolean datasets (where it holds all ti $\in$ D; $|ti| <= O(\log|I|)$) lower complexity bounds have been proved to hold. Finding rules in quantitative data further strengthens importance of efficiency. An increase in the number of values that can be associated with any given variable increases the number of rules exponentially, thus causing execution time to increase significantly [7] and [8].

2.2 Classification Techniques

Classification is the process of learning a function that maps data objects to a subset of a given class set. Therefore, a classifier is trained with a labeled set of training objects, specifying each class. There are two goals of classification:

- →Finding a good general mapping that can predict the class of so far unknown data objects with high accuracy. For this goal, the classifier is a mere function. To achieve this goal, the classifier has to decide which of the characteristics of the given training instances are typical for the complete class and which characteristics are specific for single objects in the training set.
- →The other goal of classification is to find a compact and understandable class model for each of the classes. A class model should give an explanation why the given objects belong to a certain class and what is typical for the members of a given class. The class model should be as compact as possible because the more compact a model is, the more general it is. Furthermore, small and simple class models are easier to understand and contain less distracting information. Of course, a good classifier should serve both purposes, but for most practical applications finding an accurate mapping is more important than developing understandable class models. Thus, multiple techniques are used to classify objects that do not offer an understandable class model.

Example applications of classification methods are mapping emails into one out of a determined set of folders, predicting the functional class of proteins, finding relevant information in the WWW, and predicting the costumer class for a new customer.

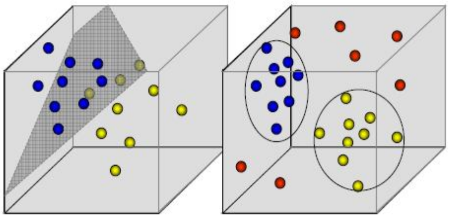

Figure 1: Classification separates the data space (left) and clustering group data objects (right).

Classification and Clustering are strongly connected. Classification tries to learn the characteristics of a given set of classes, whereas clustering finds a set of classes within a given data set. An important feature of clustering is that it is not necessary to specify a set of example objects. Therefore, clustering can be applied in applications where there is no or little prior knowledge about the groups or classes in a database. However, the usefulness of a found clustering is often subject to individual interpretation and strongly depends on the selection of a suitable similarity measure. In applications for which the existence of a dedicated set of classes is already known, the use of classification is more advisable. In these cases providing example objects for each class is usually much easier than constructing a feature space in which the predefined classes are grouped into delimited clusters. Furthermore, the performance of a classifier can easily be measured by the percentage of correct class predictions it achieves. To conclude, clustering and classification are related data mining tasks that are used in different situations. Figure 2 displays class separation by a classifier on the left side and the grouping of two clusters in a noisy data set on the right side.

The task of classification is to learn a function that maps data objects to their correct class(es) in a predefined class set. A classifier learns from a so called training set, containing a sufficient number of already mapped objects for each class. The training objects are considered to be “labeled” with the name of the class they belong to. Classification is also called supervised learning because it is directed by these labeled objects [9] and [10].

2.3 Clustering Techniques

Clustering analysis is a very broad field and the number of available methods and their variations can be overwhelming. A good introduction to numerical clustering can be found in Cluster Analysis or in Cluster Classification. A more up-to-date view of clustering in the context of data mining is available in Data Mining: Concepts and Techniques.

2.3.1 Partitioning Methods

Partitioning clustering methods divide the input data into disjoint subsets attempting to find a configuration which maximizes some optimality criterion. Because enumeration of all possible subsets of the input is usually computationally infeasible, partitioning clustering employs an iterative improvement procedure which moves objects between clusters until the optimality criterion can no longer be improved.

The most popular partitioning algorithm is the k-Means algorithm. In k-Means, we define a global objective function and iteratively move objects between partitions to optimize this function. The objective function is usually a sum of distances (or sum of squared distances) between objects and their cluster’s centers and the objective is to minimize it. The representation of a cluster can be an average of its elements (its centroid) or a mean point (object closest to the centroid of a cluster). In the latter case we call the algorithm k-Medoids. Given the number of clusters k a priori, a generic k-Means procedure is implemented in four steps:

1. Partition objects into k nonempty subsets (most often randomly),
2. Compute representation of centers for current clusters,
3. Assign each object to the closest cluster,
4. Repeat from step 2 until no more reassignments occurs.

By moving objects to their closest partition and recalculating partition’s centers in each step the method eventually converges to a stable state, which is usually a local optimum. We discuss computational complexity of k-Means in later sections, for now let us just comment that the entire procedure is efficient in practice and usually converges in just a few iterations on non-degenerated data. Another thing worth mentioning is that clusters created by k-Means are spherical with respect to the distance metric; the algorithm is known to have problems with non-convex, and in general complex, shapes.

2.3.2 Hierarchical Methods

A family of hierarchical clustering methods can be divided into agglomerative and divisive variants. Agglomerative Hierarchical Clustering (AHC) initially places each object in its own cluster and then iteratively combines the closest clusters merging their content. The clustering process is interrupted at some point, leaving a dendrogram with a hierarchy of clusters. Many variants of hierarchical methods exist, depending on the procedure of locating pairs of clusters to be merged. In the single link method, the distance between clusters is the minimum distance between any pair of elements drawn from these clusters (one from each), in the complete link it is the maximum distance and in the average link it is correspondingly an average distance (a discussion of other merging methods can be found. Each of these has a different computational complexity and runtime behavior. Single link method is known to follow “bridges” of noise and link elements in distant clusters (a chaining effect). Complete link method is computationally more demanding, but is known to produce more sensible hierarchies. Average link method is a trade-off between speed and quality and efficient algorithms for its incremental calculation exist such as in the Buckshot/ Fractionation algorithm [2] and [7].

2.3.3 Other Clustering Methods

A number of other clustering methods are known in literature; density-based methods, model-based and fuzzy clustering, self organizing maps and even biology-inspired algorithms. An interested Reader can find many surveys and books providing comprehensive information on the subject.

3. TAXONOMY OF INTRUSION DETECTION SYSTEMS

There are three main components of IDS: data collection, detection, and response. The data collection component is responsible for collection and pre-processing data tasks: transferring data to a common format, data storage, and sending data to the detection module.

IDS can use different data sources which are the inputs to the system: system logs, network packets, etc. If an IDS monitors activities on a host and detects violations on the host, it is called host-based IDS (HIDS). An IDS that monitors network packets and detects network attacks is called network-based IDS (NIDS). NIDSs generally listen in promiscuous mode to the packets in a segment of the network, allowing them to detect distributed attacks. There are also intrusion detection systems that use both host-based IDS and network-based IDS. For example, a system can use NIDS and also HIDS for important hosts in the networks such as servers, databases, and the like. Since NIDS cannot monitor encrypted packets, a hybrid approach, network node IDS (NNIDS) is introduced where each host in the network has NNIDS to monitor network packets directed to the host [3] and [4].

A data mining-based IDS is significantly more complex than a traditional system. The main cause for this is that data mining systems require large sets of data from which to train. The hope to reduce the complexity of data mining systems has led to many active research areas. First, management of both training and historical data sets is a difficult task, especially if the system handles many different kinds of data. Second, once new data has been analyzed, models need to be updated. It is impractical to update models by retraining over all available data, as retraining can take in particular time, and updated models are required immediately to ensure the protection of extended modified systems. Some mechanism is needed to adapt a model to incorporate new information. Third, many data mining-based IDSs are difficult to deploy because they need a large set of clean (i.e., not noisy) labeled training data. Typically the attacks within the data must either be manually labeled for training signature detection models, or removed for training anomaly detection models. Manually cleaning training data is expensive, especially in the context of large networks. In order to reduce the cost of deploying a system, it must be able to minimize the amount of clean data that is required by the data mining process [1] and [2] and [6].

3.1 Application of Data Mining Based Intrusion Detection

In typical applications of data mining to intrusion detection, detection models are produced off-line because the learning algorithms must process tremendous amounts of archived audit data. These models can naturally be used for off-line intrusion detection (i.e., analyzing audit data offline after intrusions have run their course). Effective intrusion detection should happen in real-time, as intrusions take place, to minimize security compromises. Now we discuss the approaches to make data mining-based ID models work efficiently for real-time intrusion detection. In contrast to off-line IDSs, a key objective of real-time IDS is to detect intrusions as early as possible. Therefore, the efficiency of the detection model is a very important consideration. Because the data mining-based models are computed using off-line data, they implicitly assume that when an event is being inspected (i.e., classified using an ID model), all activities related to the event have completed so that all features have meaningful values available for model checking.

Unfortunately, DoS attacks, which typically generate a large amount of traffic in a very short period time, are often used by intruders to first overload an IDS, and use the detection delay as a window of opportunity to quickly perform their malicious intent. For example, they can even seize control of the host on which the IDS lives, thus eliminating the effectiveness of intrusion detection altogether. It is necessary to examine the time delay associated with computing each feature in order to speed up model evaluation. The time delay of a feature includes not only the time spent for its computation, but also the time spent waiting for its readiness (i.e., when it can be computed). For example, in the case of network auditing, the total duration of a network connection can only be computed after the last packet of the connection has arrived, whereas the destination host of a connection can be obtained by checking the header of the first packet. From the perspective of cost analysis, the efficiency of an intrusion detection model is its computational cost, which is the sum of the time delay of the features used in the model [2] and [5].

3.2 Data Mining Based Intrusion Detection System Architecture

The overall system architecture is designed to support a data mining-based IDS with the properties described throughout now. As shown in Figure 2, the architecture consists of sensors, detectors, a data warehouse, and a model generation component. This architecture is capable of supporting not only data gathering, sharing, and analysis, but also data archiving and model generation and distribution. The system is designed to be independent of the sensor data format and model representation. A piece of sensor data can contain an arbitrary number of features. Each feature can be continuous or discrete, numerical or symbolic. In this framework, a model can be anything from a neural network, to a set of rules, to a probabilistic model. To deal with this heterogeneity, an XML encoding is used so each component can easily exchange data and/or models. The design was influenced by the work in standardizing the message formats and protocols for IDS communication and collaboration: the Common Intrusion Detection Framework and the more recent Intrusion Detection Message Exchange Format, by the Intrusion Detection Working Group of IETF, the Internet Engineering Task Force). Using CIDF or IDMEF, IDSs can securely exchange attack information, encoded in the standard formats, to collaboratively detect distributed intrusions.

In the architecture, data and model exchanged between the components are encoded in the standard message format, which can be trivially mapped to either CIDF or IDMEF formats. The key advantage of the architecture is its high performance and scalability. That is, all components can reside in the same local network, in which case, the work load is distributed among the components; or the components can be in different networks, in which case, they can also participate in the collaboration with other IDSs in the Internet.

3.2.1 Sensors

Sensors observe raw data on a monitored system and compute features for use in model evaluation. Sensors insulate the rest of the IDS from the specific low level properties of the target system being monitored. This is done by having the entire sensors implement a Basic Auditing Module (BAM) framework. In a BAM, features are computed from the raw data and encoded in XML.

3.2.2 Detectors

Detectors take processed data from sensors and use a detection model to evaluate the data and determine if it is an attack. The detectors also send back the result to the data warehouse for further analysis and report. There can be several (or multiple layers of) detectors monitoring the same system. For example, workloads can be distributed to different detectors to analyze events in parallel. There can also be a “back-end” detector, which employs very sophisticated models for correlation or trend analysis, and several “front-end” detectors that perform quick and simple intrusion detection. The front-end detectors keep up with high-speed and high-volume traffic, and must pass data to the back-end detector to perform more thorough and time consuming analysis.

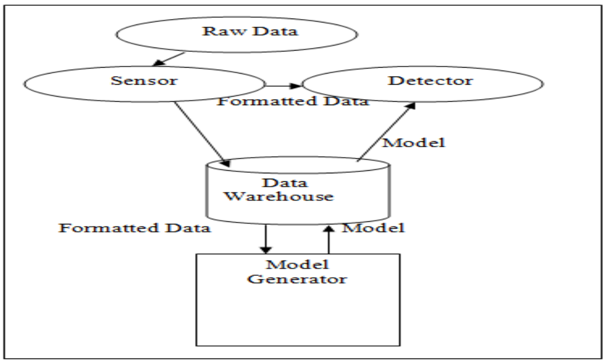

Figure 2: The Architecture of Data Mining based IDS

3.2.3 Data Warehouse

The data warehouse serves as a centralized storage for data and models. One advantage of a centralized repository for the data is that different components can manipulate the same piece of data asynchronously with the existence of a database, such as off-line training and manually labeling. The same type of components, such as multiple sensors, can manipulate data concurrently. Relational database features support “stored procedure calls” which enable easy implementation of complicated calculations, such as efficient data sampling carried out automatically on the server. Arbitrary amount of sensor data can also be retrieved by a single SQL query. Distribution of detection models can be configured to push or pull.

The data warehouse also facilitates the integration of data from multiple sensors. By correlating data/results from different IDSs or data collected over a longer period of time, the detection of complicated and large scale attacks becomes possible.

3.2.4 Model Generator

The main purpose of the model generator is to facilitate the rapid development and distribution of new (or updated) intrusion detection models. In this architecture, an attack detected first as an anomaly may have its exemplary data processed by the model generator, which in turn, using the archived normal and intrusion data sets from the data warehouse, automatically generates a model that can detect the new intrusion and distributes it to the detectors (or any other IDSs that may use these models). Especially useful are unsupervised anomaly detection algorithms because they can operate on unlabeled data which can be directly collected by the sensors.

A prototype implementation of a data mining and CIDF based IDS. In this system, a data mining engine, equipped with feature extraction programs and machine learning programs, serves as the model generator for several detectors. It receives audit data for anomalous events (encoded as a GIDO, the Generalized Intrusion Detection Objects) from a detector, computes patterns from the data, compares them with historical normal patterns to identify the “unique” intrusion patterns, and constructs features accordingly. Machine learning algorithms are then applied to compute the detection model, which is encoded as a GIDO and sent to all the detectors. Much of the design and implementation efforts had been on extending the Common Intrusion Specification Language (CISL) to represent intrusion detection models. The preliminary experiments show that the model generator is able to produce and distribute new effective models upon receiving audit data [1] and [3] and [10].

4. RELATED WORK

Rakesh Shrestha et al. proposed a novel cross layer intrusion detection system in MANET. This paper, descried a novel cross layer intrusion detection architecture to discover the malicious nodes and different types of DoS attacks by exploiting the information available across different layers of protocol stack in order to improve the accuracy of detection. They used cooperative anomaly intrusion detection with data mining technique to enhance the proposed architecture. They implemented fixed width clustering algorithm for efficient detection of the anomalies in the MANET traffic and also generated different types of attacks in the network. The simulation of the proposed architecture is performed in OPNET. In this paper gives brief description about cross layer techniques in IDS followed by association module.

Association Module

Once association rules are extracted from multiple segments of a training data set, they are then aggregated into a rule set. The feature sets consist of control and data frames from MAC frames and control packets like RREQ, RREP and RERR including data packets of IP packets from network layer. All the control packets are combined into one category as routing control packet and IP data packet as routing data packet. So, the payloads in MAC data frames contain either a routing CtrlPkt or routing DataPkt. The feature set is foreshortened by associating one or more features from different layers to specific MAC layer feature so that the overhead of learning is minimized. The characteristics are assorted based on dependency on time, traffic.

They find all non-empty subsets of f to generate rules for every frequent itemset f. For every subset a, we output a rule of the form a=> (f-a) if the ratio of support (f) to support (a) is atleast minconf. All subsets of f are considered to generate rules with multiple consequents.

The authors used data mining techniques in Intrusion detection module in order to improve the efficiency and effectiveness of the MANET nodes. With the studies, found out that among all the data mining intrusion detection techniques, clustering-based intrusion detection is the most potential one because of its ability to detect new attacks. Many traditional intrusion detection techniques are limited with collection of training data from real networks and manually labeled as normal or abnormal. It is very time consuming and expensive to manually collect pure normal data and classify data in wireless networks. They used association algorithm such as Apriori which can be utilized to achieve traffic features which is then followed by clustering algorithm.

The association rule and clustering are used as the root for accompanying anomaly detection of routing and other attacks in MANET shown figure 3.

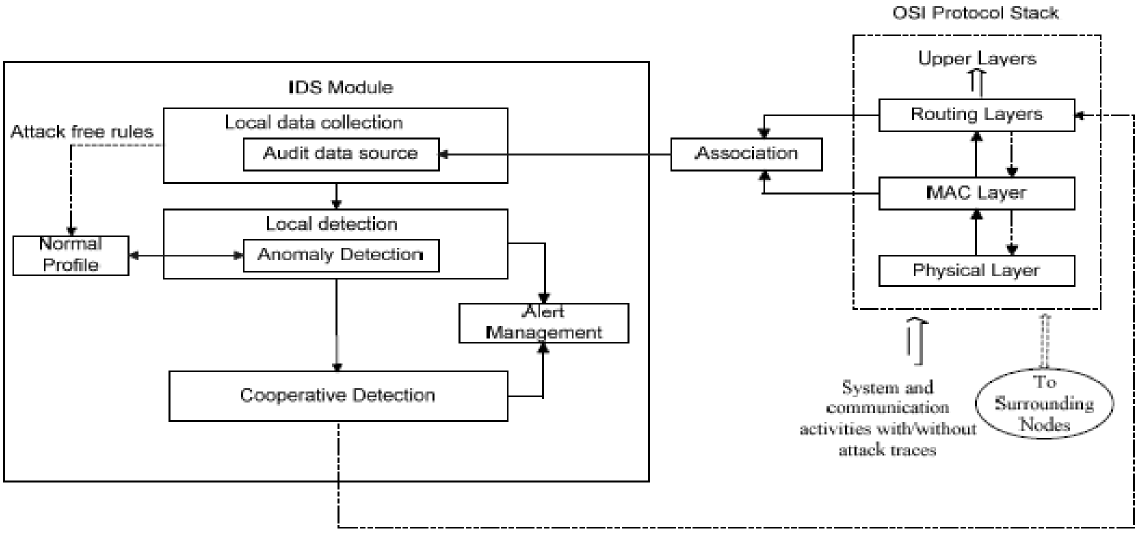

Figure 3: Association Rule based IDS Architecture in MANET

Hence, a better intrusion detection mechanism based on anomaly detection is presented in this paper utilizing cluster data mining technique. We have implemented the proposed architecture with fixed width algorithm and done the simulation and analyzed the result. Our proposed cross-layer based intrusion detection architecture is designed to detect DoS attacks and sink hole attack at different layers of the protocol stack. It is able to detect various types of UDP flooding attack and sink hole attack in an efficient way. Future work involved research into more robust and intelligent IDS system which includes further analysis of the simulation results with richer semantic information [1].

Preetee K. Karmore et al proposed detecting intrusion on AODV based Mobile Ad Hoc Networks by k-means clustering method of data mining. They implemented k-means clustering algorithm of data mining for efficient detection of intrusions in the MANET traffic and also generated black hole attacks in the network. In data mining, clustering is the most important unsupervised learning process used to find the structures or patterns in a collection of unlabeled data. Used the K-means algorithm to cluster and analyze. Data Mining Technique to detect intrusion Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining can be divided into four types: association analysis, sequence analysis, classification analysis and cluster analysis. Classification algorithm about Data Mining can be used to construct classifier, after the invasion of a large number of data sets being trained. Classifier can be used for intrusion detection.

Clustering analysis algorithm can be used to construct the network model of normal behavior, or intrusion behavior model. Association analysis algorithm can be used to describe the invasion of behavior patterns of association rules, through these rules intrusion detection can come. They used clustering method of data mining to detect an intrusion so that security of MANET will be improved.

K-means clustering is the partitioning method of data mining. They used this k-means algorithm for constructing the clusters of data. In proposed system, k-means algorithm is used to construct the centroids of clusters. This IDS system is host based which monitor each and every node in the network whether any node in the network generates any events or not. If any, then the features of that node is extracted, calculates the mean square error and then check Euclidean distance from the centroids which have been constructed previously. If it is nearest to normal cluster centroid then IDS will assume that the node is normal and it will allow preceding its events normally, if it is nearest to abnormal or malicious cluster centroid then it will not allow proceeding that is the IDS will drop an event from the queue which is generated by malicious node. In this way it is detected malicious nodes in mobile ad-hoc networks and avoided the effect of it. They used data mining technique in order to improve the efficiency and effectiveness of the mobile ad-hoc network nodes. Intrusion detection mechanism is presented in this paper utilizing cluster data mining technique. They implemented the architecture with K-means clustering algorithm and done the Simulation and analyzed the result. Their proposed intrusion detection architecture is designed to detect black hole attack [2]. In future it mechanism could be more reliable in wireless MANET environment using extended intrusion detection method.

Shaik Akbar et al. proposed Intrusion Detection System Methodologies Based on Data Analysis. In this paper focused on detailed study of different types of attacks using in KDD99CUP Data Set and classification of IDS are also presented. They are Anomaly Detection System, Misuse Detection Systems. Different Data Analysis Methodologies also explained for IDS. To identify eleven data computing techniques associated with IDS are divided groups into categories. Some of those methods are based on computation such as Fuzzy logic and Bayesian networks, some are Artificial Intelligence such as Expert Systems, agents and neural networks some other are biological concepts such as Genetics and Immune systems. In future would like to investigate the efficient technique for feature reduction of the input dataset and find out how fuzzy logic, data mining, genetic algorithms along with neural networks can help to improve intrusion detection and most of all anomaly detection [3].

Abhinav Srivastava et al. proposed Database Intrusion Detection using Weighted Sequence Mining. They proposed an algorithm for finding dependencies among important data items in a relational database management system. Any transaction that does not follow these dependency rules are identified as malicious. They show that this algorithm can detect modification of sensitive attributes quite accurately. They also suggest an extension to the Entity-Relationship (E-R) model to syntactically capture the sensitivity levels of the attributes. These approaches for database intrusion detection are using a data mining technique which takes the sensitivity of the attributes into consideration in the form of weights. Sensitivity of an attribute signifies how important the attribute is, for tracking against malicious modifications. This approach mines dependency among attributes in a database. The transactions that do not follow these dependencies are marked as malicious transactions. Divide the work of mining weighted data dependency rules for database intrusion detection into the following three components:

- Security Sensitive Sequence Mining
- Read-Write Sequence Generation
- Weighted Data Dependency Rule Generation

The proposed database intrusion detection system generates more rules as compared to non-weighted approach. There is a need for a mechanism to find out which of these new rules are useful for detecting malicious transactions. Such a mechanism helps in discarding redundant rules. They plan to use some learning mechanism to filter out extra rules generated by their approach [5].

N. Pratik Neelakantan et al. proposed IDS based a novel hybrid model that efficiently selects the optimal set of features in order to detect 802.11-specific intrusions. This model for feature selection uses the information gain ratio measure as a means to compute the relevance of each feature and the k-means classifier to select the optimal set of MAC layer features that can improve the accuracy of intrusion detection systems while reducing the learning time of their learning algorithm. Select the best set of MAC layer features that efficiently characterize normal traffic and distinguish it from abnormal traffic containing intrusions specific to wireless networks. Our framework uses a hybrid approach for feature selection that combines the filter and wrapper models. In this approach, rank the features using an independent measure: the information gain ratio. The k-means classifier’s predictive accuracy is used to reach an optimal set of features which maximize the detection accuracy of the wireless attacks. To train the classifier, the first collect network traffic containing the wireless intrusions, namely, the de-authentication, duration, and fragmentation. This paper presenting an approach to select the best features for detecting intrusions in 802.11- based networks. This approach is based on a model which combines the filter and wrapper models for selecting relevant features. In future planning to do a comparative study of the impact of the reduced feature set on the performance of classifiers based MANETs [4].

S. P. Manikandan et al. proposed Evaluation of Intrusion Detection Algorithms for Interoperability Gateways in Ad Hoc Networks. In this paper they evaluate Intrusion Detection Classification Techniques on a MANET gateway connecting to a wired network. Bayesian and Decision tree induction techniques are evaluated and results presented.

They investigate the classification efficiency of various classification algorithms for Network intrusion detection at the network gateway of the proposed MANET. The proposed MANET consists of nodes which dynamically enter the network or leave the network and move randomly within the network. The network has a single gateway connecting to the internet cloud. Nodes in the network can access the internet by establishing a communication link with the gateway. The connection between the source and the gateway can either be single hop or multi-hop.

Classification Accuracy Measurements

The TCP dump containing HTTP data from the KDD 99 dataset is used to evaluate our methodology. The dump consists of normal data along with attacks including port sweep, ip sweep, backdoor and Neptune. 0.875 % of the available dump contained abnormal packets. The attributes captured for the classification includes status of the connection, number of bytes from source to destination, number of bytes from destination to source, number of compromised conditions, number of connections on access control files and traffic features.

Information gains are used to preprocess the data, specifically for data reduction. Information gain provides the effectiveness of an attribute in classifying the training data based on the attribute which has to be predicted also called as the class label, the information gain is computed. Two different large datasets consisting of 254032 packets were used in this work with one dataset containing very few anomalous data and the other set containing more anomalous data. As MANET's operate in a power constrained environment, reducing the computational time is essential and hence only 65% of the attributes available were used. Information gain was used for feature selection. It is also found that decision tree based algorithms were able to consistently classify intrusion data as against Naïve Bayesian method. Further investigation needs to be done on effect of malicious node in the performance of an Ad hoc Network [6].

5. CONCLUSION

Data mining is the process of analyzing data from different perspectives and summarizing into useful information. Data mining can be divided into four types: association analysis, sequence analysis, classification analysis and cluster analysis. Classification algorithm about data mining can be used to construct classifier, after the invasion of a large number of data sets being trained.

The use of mobile ad hoc networks (MANETs) has increased the requirement of security in MANETs. Due to the vulnerability of ad hoc networks, intrusion prevention measures such as encryption and authentication are not enough, therefore, there is a strong need for intrusion detection as a frontline security research area for ad hoc network security.

An intrusion detection system aims to detect attacks on mobile nodes or intrusions into the networks. In this survey paper, we tried to inspect the data mining based intrusion detection system, which may be a main disturbance to the operation of it. We then discussed some typical vulnerability on data mining based intrusion detection in the mobile ad hoc networks, most of which are caused by the characteristics of the mining based secure mobile ad hoc networks such as intrusion detection, association rule, classification techniques, clustering algorithm, constantly changing topology, open media. Finally, we introduce the security schemes in the mobile ad hoc networks that can help to protect the mobile ad hoc networks.

During the survey, we also find some points that can be further explored in future, such as finding some effective security solutions and protecting the data mining based MANET using modified clustering technique and detection of Black hole routing attack with efficient intrusion detection mechanism. We can also explore much more in this research area.

Table1. After reviewing different techniques we define the Merits and Demerits of techniques

Techniques	Merits	Demerits
Novel IDS	Cross-layer based intrusion detection architecture is designed to detect DoS attacks and sink hole attack at different layers of the protocol stack. It is used to detect various types of UDP flooding attack and sink hole attack in an efficient way.	This system is not reliable and intelligent IDS system which needs further analysis of the simulation results with richer semantic information [1].
Intrusion Detection System, k-means clustering algorithm	The intrusion detection architecture is designed to detect black hole attack. The aim is to improve the detection rate and decrease the false alarm rate.	The whole system does not provide sufficient mining method and security from more active attacks that a malicious node can perform against the routing protocol [2]
Data Dependency, Weighted Sequence Mining, Intrusion Detection	The proposed database intrusion detection system generates more rules as compared to non-weighted approach. There is a need for a mechanism to find out which of these new rules are useful for detecting malicious transactions. Such a mechanism helps in discarding redundant rules. It plans to use some learning mechanism to filter out extra rules generated by this approach.	The Sequence number attack and detect system could be extended to operate for proactive routing protocols like DSDV. The main architecture and the high-level components of the system will remain the same, the only thing that changes are the patterns that signify the attacks, and of course, the threshold values that should be modified to match the implementation of the underlying protocol [5].

Hybrid IDS	Approach is based on a model which combines the filter and wrapper models for selecting relevant features.	The system can also be extended to include some cryptographic mechanism like a certification authority that would prevent nodes from impersonating other nodes. This would provide a more complete security solution however it would introduce additional overhead [4].
KDD, Anomaly Detection System	This paper is a brief survey and useful for all researchers who want to investigate more efficient methods against intrusions.	This paper needs to survey more recent techniques for efficient knowledge collection [3].

References

[1] Rakesh Shrestha, Kyong-Heon Han, Dong-You Choi, Seung-Jo Han, “A Novel Cross Layer Intrusion Detection System in MANET”, 2010 24th IEEE International Conference on Advanced Information Networking and Applications, pp 647-656.

[2] Preetee K. Karmore , Smita M. Nirkhi, “Detecting Intrusion on AODV based Mobile Ad Hoc Networks by k-means Clustering method of Data Mining”, International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, pp 1774-1779.

[3] Shaik Akbar, Dr.K.Nageswara Rao and Dr.J.A.Chandulal, “Intrusion Detection System Methodologies Based on Data Analysis”, International Journal of Computer Applications (0975 – 8887) Volume 5– No.2, August 2010, pp 10-20.

[4] N. Pratik Neelakantan, C. Nagesh, “Role of Feature Selection in Intrusion Detection Systems for 802.11 Networks”, International Journal of Smart Sensors and Ad Hoc Networks (IJSSAN) Volume-1, Issue-1, 2011, pp 98-101.

[5] Abhinav Srivastava, Shamik Sural and A.K. Majumdar, “Database Intrusion Detection using Weighted Sequence Mining”, JOURNAL OF COMPUTERS, VOL. 1, NO. 4, JULY 2006, pp 8-17.

[6] S. P. Manikandan and Dr. R. Manimegalai, “Evaluation of Intrusion Detection Algorithms for Interoperability Gateways in Ad Hoc Networks”, International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 9 September 2011, pp 3243-3249.

[7] Prakash Ranganathan, Juan Li, Kendall Nygard, “A Multiagent System using Associate Rule Mining (ARM), a collaborative filtering approach”, IEEE 2010, pp- v7 574- 578.

[8] Prof Thivakaran.T.K, Rajesh.N, Yamuna.P, Prem Kumar.G, “PROBABLE SEQUENCE DETERMINATION USING INCREMENTAL ASSOCIATION RULE MINING AND TRANSACTION CLUSTERING”, IEEE 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, pp 37-41.

[9] Shuang Deng and Hong Peng, “Document Classification Based on Support Vector Machine Using A Concept Vector Model”, Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.

[10] Choochart Haruechaiyasak, Mei-Ling Shyu and Shu-Ching Chen, “Web Document Classification Based on Fuzzy Association”, Proceedings of the 26 th Annual International Computer Software and Applications Conference (COMPSAC’02).