Survey on Intrusion Detection Using Data Mining Methods

Lokendra Singh Parihar[1], Akhilesh Tiwari[2]
[1, 2] Department of CSE & IT
[1, 2] Madhav Institute of Technology and Science, Gwalior, MP. (India) - 474005

Abstract- *The present emerging information growth made so numerous challenges in the data mining. Data mining is the procedure of removing valid, before known & comprehensive datasets for the future decision making. As the better technology through WWW the streaming information come into picture with its challenges. The data which alteration with time & update its value is known as streaming information. At the most of the data is streaming in nature, there are numerous challenges need to face in the security perspective sense. IDS works in the detecting supposition the intruders to protect the respective method. The research in mining of data stream & system of Intrusion detection gained high attraction because of the system's safety measure significance. Algorithms, frameworks & systems that address security issue have been developed over few years. An current System of Intrusion Detection needs detection rate and high accuracy as well as low false alarm rate. We briefly define and compare a big amount of intrusion detection methods, techniques and systems. In addition, we also discuss tools which are used through network defenders and datasets.*

Keywords:- Intrusion Detecting System, Data Mining, Clustering, Attacks.

I. INTRODUCTION

Data mining is the withdrawal of unseen predictive data or information from a big amount of database. It is strong and novel technology has great prospective to companies focus on the most significant information in their information repository. Data mining tools predict future drift and behaviors through permitting businesses to make knowledge-dive decisions [1].

Data mining mechanism can answer business or profession questions which were classically taking a big amount of time consuming to resolve. In the traditional data set, data does not change with time and they are nature is static, whereas streaming information generated continuously. Continuous data, i.e. streaming data is impossible to store, hence it need to be analyzed in single pass [2] [3] [4].Streaming data can be network data which consists of inbound and outbound traffic of the network.

With the development of network technology, now a days more and more people learn various ways of attack through the rich network resources, and carry out extremely destructive attack through simple operation. In recent years, the amount of hackers' attack is growing 10 times per year. Therefore, it has become the urgent topic to ensure the computer systems, network systems as well as the entire information infrastructure security, and it has become the general concern of the computer industry that how to detectand prevent these attacks effectively.

There are many methods to strengthen the security of network at moment, for example encryption, VPN, firewall, etc., but all of these are too static to provide an efficient protection.

However, intrusion detection is a dynamic one,which can gives dynamic protection to the network security in monitoring, attack and counter-attack.

II. DATA MINING AND INTRUSION DETECTION

a) Data Mining Technology

Data Mining means extracting or mining the knowledge from the mass data. To be specific, it means processing data so as to gain the implied, prior unknown, potential and useful knowledge, which can be expressed as patterns. The targets of digging include not only data source and file system, but also any data collections such as Web source[5]. Data mining is the newest presented intrusion detection methodology. Its benefit lies in the fact that it can withdraw the needed and unknown information and regularities from the massive network information and host log data. It is a novel attempt to use data mining in achieving network security, both at home and abroad.

At the moment, the research on data mining algorithm is quite mature, and data mining itself is a general knowledge discovering technique. In the field of intrusion detection, we consider it as a data analysis process, which applies certain data mining algorithm to massive safe data in order to construct system of intrusion detection with selfadaptability and scalability. At current, an algorithm of data mining applied to intrusion detection mostly has four basic patterns: association, clustering, sequence and classification.

b) Intrusion Detection Technology

Security had become major concern in all fields of network & system infrastructure[6]. The basic challenge is to authorized user identify & the one who is legitimate to system access without abusing their privileges. Insider threats as well as outsider threats are rigorous to the system/network, known as intruders. Intrusion detection methodology can be describe as a method that classifies and deals with the malicious use of network and computer resources. It contain the exterior method behavior of intrusion and internal user's non-authorized. It is a methodology designed to ensure the computer system security that can discover and inform the non-authorized and abnormal occasions, used to detect the violation of network security.

An Intrusion Detection System (IDS) is critical technology to detect such intruders who are system harmful. Basic aim of the IDS is to protect the system & network from the intruders. IDS keep track of activities behavior; if they are system malicious then it'll be automatically detected through the IDS [7].

Thus IDS is further classified into three categories as follows[8]:

i) NIDS It is an platform independent that classifies intrusions through examining traffic network and monitors numerous hosts. NIDS increase access to network traffic through network hub connecting, port mirroring network switch configured, or network tap.

ii) HIDS It consists of an agent on a host that classifies intrusions through system calls analyzing, application logs, modifications of file-system (Access control lists, binaries, password files, capability databases, etc.) and other different host state and activities. In a HIDS, sensors commonly consist of a software agent.

iii) Hybrid IDS It complements HIDS system through the ability the monitoring network traffic for a particular host; it is various from the NIDS that monitors all network traffic . In computer security, a NIDS is an intrusion detection system that attempts to discover unauthorized access to a computer network through analyzing traffic on the network for signs of malicious activity.

In the case of detecting data target, intrusion detecting system can be classified as host-based, network-based, kernel-based and application-based. In this thesis, we focuson the construction of network -based system.

According to the differences of data analysis methods (that is, detection methods), we can say that there are two types of intrusion detecting system.

1) Misuse detection

Misuse Detection refers to the confirming attack events through matching features by attacking feature library. It advances in the detection high speed and false alarm low percentage. However, it fails in non-pre-designated attacks discovering in the service library, so it cannot discover the many novel attacks.

2) Anomaly detection

Anomaly detection refers to features of storing consumer's normal actions into database, then comparing consumer's present behavior with those in the database. If the divergence is big enough, we can say that there is something abnormal. Its merits lie in its comparative irrelevance with system, its strong versatility and detect possibility the attack that has never been detected. But because of the fact that contour conducted cannot give a full fill description of each users'in the system behaviors, moreover allconsumer's behavior modifications constantly, its main disadvantage is the false alarm at high rate.

Combining these two, we may obtain a better performance. Anomaly detection can detect new, unknown attack or likewise, while misuse detection prevent the occasion that patient hacker gradually change pattern behavior so as to make the anomaly detection legalize the attack, which protects the integrity of anomaly detection. Intrusion Detection knowledge sources can be obtained with some dedicated capture tool. In Windows, data packs aregained with Wincap; in Unix, with Tcpdump and Arpwatch.

c) Need of Data Mining In Intrusion Detection

Data Mining refers to the removing hidden procedure, previously unknown and valuable knowledge from big amount databases. It is a convenient extracting patterns way and on issues focuses relating to the their feasibility, efficacy, scalability and efficiency. Thus, data mining method help to detect patterns in the set of data and also use patterns to the detect future intrusions in same data. The following are a some particular things that create the use of data mining significant in system of intrusion detection:

i) Manage rules of firewall for anomaly detection.
ii) Analyze network data big volumes.
iii) Similar data mining tool can be applied to various data sources.
iv) Achieves data visualization and summarization.
v) Differentiates information that can be used for analysis of deviation.
vi) Clusters the introduction into groups such that it possess high intra-class similarity and also low inter-class similarity.

Data Mining Techniques for Intrusion Detection Systems

Data Mining refers to the extracting hidden procedure, previously unknown and valuable information from large databases. It is a convenient way of extracting forms and focuses on problems relating to their feasibility, efficiency, utility and scalability. Thus data mining procedures help to detect forms in the data set and use these forms to detect future intrusions in related data. The following are a few specific things that create the usage of data mining significantin an IDS:

i) Arrange firewall rules for anomaly detection.

ii) Analyze large volumes of network data.

iii) Similar data mining tool can be applied to dissimilar data sources.

iv) Performs data summarization and visualization.

v) Dissimilar data that can be used for deviance analysis.

vi) Clusters the data into groups such that it possess great intra-class similarity and low inter-class similarity.

Data Mining Techniques for Intrusion Detection Systems

Data mining methods perform significant role in IDS. Various data mining methods for example association rule mining, classification, clustering are used normally to acquire information about intrusions by observing and analyzing the network data. The following defines the different data mining techniques:

A. Classification: It is a managed learning technique. A classification based IDS will classify all the network traffic into either common or malicious. Classification technique is normally used for anomaly detection. The classification process is as follows:

i) It receives collection of item as input.
ii) Maps the items into predefined groups or classes define by specific qualities.
iii) After mapping, it outputs a classifier that can correctly predict the class to which a fresh item belongs.

B. Association Rule: This procedure searches a frequently occurring item set from a large dataset. Association rule mining determines association rules and/or correlation relationships amongst large set of data items. The mining procedure of association rule can be distributed into two steps as follows:

i) Frequent Item set Generation Creates all set of items whose support is better than the identified threshold called as minsupport.
ii) Association Rule Generation From the earlier created frequent item sets, it creates the association rules in the form of ― at that time‖ statements that have confidence better than the identified threshold known as min confidence.

The basic steps for incorporating association rule for intrusion detection are as follows:

i) The network data is settled into a database table where all row signifies an audit record and each column is a field of the audit records.
ii) The intrusions and user activities shows frequent connections among the network data. Consistent network data behaviors can be captured in association rules.
iii) Rules based on network data can constantly merge the rules from a fresh run to aggregate rule set of totally previous runs.
iv) Thus with the association rule, we get the ability to capture behavior for properly detecting intrusions and hence lowering the false alarm rate.

C. Clustering: It is an unverified machine learning mechanism for discovering forms in unlabeled data. It is used to label information and allocate it into clusters where all cluster comprises of members that are quite similar. Members from dissimilar clusters are different from all other. Hence clustering techniques can be valuable for classifying network data for detecting intrusions. Clustering can be applied on both Misuse detection and Anomaly detection. The mainlevel involved in classifying intrusion are follows :

i) Find the biggest cluster, which involves of maximum number of orders and label it as common.
ii) Sort the remaining clusters in an ascending order of their spaces to the biggest cluster.
iii) choose the first K1 clusters so that the various introduction instances in these clusters sum up to the

¼`N and label them as common, where ` is the percentage of common instances.

iv) Label each other clusters as malicious.

v) After clustering, heuristics are used to automatically label all cluster as also common or malicious. The self- labeled clusters are then used to detect attacks in a single test dataset.

III. APPLICATION OF DATA MINING IN INTRUSION DETECTION

In traditional intrusion detecting system, security experts firstly classify attacking actions and system weakness, choose statistical methods due to the detecting types, then manually enter the code and establish the corresponding detecting rules and modes. For complex network system, the limitation of experts' knowledge grows with the change of time and space, so it is not good to improve the effectiveness of detecting the intrusion detecting modes.Security experts usually concern about the known attacking features and system weakness and research on that,which causes the lack of adaptability of the detecting patternto the unknown intrusion that the system is about to befacing. Meanwhile, the long upgrade cycle of security system, the high cost, these are not advantageous for improving the adaptability of intrusion detecting pattern.

As the experts' rules and statistical methods often requirethe support of software and hardware, it stops the system from reusing and developing in new environment,meanwhile it causes the difficulty of embedding new detecting modules. All of these are not good for improving scalability of intrusion detecting pattern.Therefore, it has become an important issue how to establish an effective, self-adaptable and scalable intrusion detecting pattern in intrusion detecting field. Considering intrusion detection as a data analysis procedure through applying the data mining predominance in its effective use of knowledge, this is a technique that can automatically create accurate and applicable intrusion patterns from massive audit data, which creates intrusion detecting system can be appliedto any computer environment. This approach has become apopular topic of research, in the field of inter discipline of network security and artificial intelligence. The analysis methods of association, sequence, clustering and classification in the mining of data has been proved possible.

Drawback of current IDS:

- Current IDS does not detect the novel intruders: As some of the IDS work on the signature based technology, there are some predefined signatures in IDS, but as the signatures are predefined they fail to detect the novel intruders.
- False Positive: It occurs when normal is wronglyclassified as intruder.
- False Negative: It occurs when an intruder is wrongly classified as normal.

IV. CLASSIFICATION TECHNIQUES

Classification is a mining of information function that allocates items in a set to goal classes or categories. The classification goal is to precisely predict the goal class for all case in the information. [5][6]

- Association rule mining
- Bayesian Classification
- Decision tree classification
- Nearest Neighbor
- NN (Back Propagation)
- SVMs

1. Association Rule Mining

Association rule mining, one of the advance wide and well researched data mining methods.[31] It goals to mine uncommon relationship, associations, frequent patterns or casual structures among items sets in the transaction databases or other different image repositories. Association rules are extensively used in numerous areas for example telecommunication networks, risk management and market, inventory control etc. Numerous association mining methods and algorithms will be briefly compared and introduced later. Association rule mining is to discover out association rules that fulfill the predefined minimum confidence and support from a provide database. The problem is generally decomposed into sub problems. One is to find those item sets whose happenings exceed a predefined threshold in the database; those item sets are known as large or frequent item sets. The second issue is to create association rules from those huge item sets with the minimal confidence constraints.

2. Bayesian Classification

A naive Bayes classifier assumes that the absence or presence of a specific feature is unrelated to the absence or presence of any other different feature, provide the class variable. A naive Bayes classifier reflects features to contribute independently to the probability that this fruit is an apple, regardless of the absence or presence of the other different features. Naive Bayes classifiers can be trained more effectively in a supervised learning setting. A naive Bayes benefit is that it only needs a less quantity of training information to estimate the parameters (variances and means of the variables) required for classification. Because independent variables are assumed, only the variances of the variables for all class requirement to be determined and not the complete covariance matrix.

3. Decision Tree Classification

Decision tree classification method is most valuable in the classification issue. It is a flow chart for example tree structure. Trees are created in a top down recursive divide and conquer manner. In this classification technique used in various kind algorithm to categorize the introduction sets, the algorithms are: [29]

- ID3 (Iterative Dichotomiser)
- C4.5 (a Successor of ID3)
- CART (Classification and Regression Trees)

The algorithm follows a top-down method, which starts with a training tuples set and their associated class labels.

Advantages: Rules can be created that are simply to understand and interpret. It is scalable for big database because the size of tree is database size independent. [31] All tuple in the database must be filtered by the tree, and time is proportional to the tree height.

Disadvantages: It is does not handle continuous information. [31]Handling missing information is difficult because correct branches in tree could not be taken the labels.

4. Nearest Neighbor

Nearest neighbor classifiers are based on learning through analogy. The training samples are defined through n dimensional numeric attributes. All sample represents a point in an n-D space. In this way, each of the training samples are stored in the space of an n-dimensional of pattern. When provide an unknown sample, a k-nearest neighbor classifier searches the space pattern for the k training samples that are closest to the unknown sample. "Closeness" is well-defined in Euclidean distance terms. The unknown sample is allocated the most general class among its k nearest neighbors. When k=1, the unknown sample is allocated the training sample class that is closest to it in space of pattern. Nearest neighbor classifiers are lazy learners and instance-based in that they store each of the training samples and do not construct a classifier until a novel (unlabeled) sample requirements to be classified.[33]

5. Neural Networks

NN show an image of brain or symbol for knowledge processing. [29] [31] These models are biologically stimulated rather than an exact replica of how the brain essentially functions. NN have been present to be most talented systems in numerous forecasting business and applications classification applications because of their ability to “learn” from the information, their non parametric nature (i.e., no rigid assumptions), and their facility to generalize. Neural computing refers to a methodology of recognition of pattern for machine learning. The model from neural computing is often known as ANN or a NN. NN have been used in numerous business applications for pattern recognition, classification, forecasting, and prediction.

6. Support Vector Machine

SVM has developed as one of the general and valuable methods for information classification [32]. It can be used for classify the both non linear and linear information. The SVM objective is to create a model that predicts the aim information value incidence in the testing set in which only attributes are provide. The classification aim in SVM is to separate the two different classes through function means prepare from presented information and thereby to create a classifier that will work well on further unseen information. The simplest SVM form classification is the classifier of maximal margin. It is used to resolve the most common classification issue, namely the binary classification case with linear separable training information. [29] The maximal margin goal classifier is to discover the hyperplane with the main margin, i.e., the maximal hyperplane, in real-world issue, training information are not always linear separable. In order to the handle nonlinearly separable cases few slack variables have been presented to SVM so as to tolerate few training errors, with noise influence in training information thereby reduced. This classifier with slack variables is referred to as a soft-margin classifier.

V. RELATED WORK

G. V. Nadianmai and M. Hemalathain [9] in their paper ―Effective approach toward Intrusion Detection System using data mining techniques‖, considered four issues namely Classification of Data, Human Interaction High Level, Labeled Data Lack, and DDOS attack Effectiveness and solved them applying the proposed algorithms of EDADT, Hybrid IDS model, Semi-Supervised Method and Varying HOPE RAA algorithm respectively. To solve the problem related to data classification, an enhanced data adapted decision tree algorithm is implemented which effectively data

data classifies into normal and attack without any classification. To minimize the network administrator workload, a human interaction high level based on SNORT and anomaly based approaches are being used. This has a Hybrid IDS that automatically categorizes the data based on the pre-defined rules within it. The issue related to belling the unlabeled data is solved applying Semi-Supervised Method where with the less quantity of labeled data, the unlabeled data big amount can be labeled. The last problem related to Distributed Denial of Service Attack is addressed by applying varying clock drift. This varying clock drift in network based applications makes it difficult for the intruder to access the port that has been used through the legitimate client.

W. Feng et al. [10] in their paper ―Mining network data for intrusion detection through combining SVMs with Ant Colony Networks‖ achieved better performance in both faster running time and detection accuracy rate through combining two different existing machine learning approaches (SVM and CSOACN). Their proposed work is based on five main interactive modules. The basic contributions of this paper conclude the modifications to the learning of supervised SVM and the learning of unsupervised CSOACN so that they can be used together efficiently and interactively. It also combines the modified CSOACN and SVM to minimize the training data set while permittingnovel data points to be added to the training set.

The number of selected records from each difficulty level group is inversely proportional to the records percentage in the original KDD data set[11]. As a result, the classification rates of distinct machine learning approaches vary in a wider range, which creates it extraeffective to contain an accurate evaluation of various learning methods. The numbers of records in the train and test sets are reasonable, which creates it affordable to run the experiments on the complete set without the necessity to randomly select a less portion. Consequently, evaluation outcomes of various research works will be consistent and comparable.

The proposed work by Karim Al-Saedi et al. [12] in their paper ―Research Proposal: An Intrusion Detection Alert Reduction System and also Data Mining based assessment Framework is IDS ARADMF which holds three different systems: Traffic data retrieval and also system of collection mechanism, reduction IDS alert procedures system IDS ready framework process. The movement information recovery and gathering component frameworks add to a system to spare IDS alarms, remove the standard elements as interruption location trade configuration and spare them in DB file(CSV-type). It include the IDMEF which works as procurement alerts and field reduction is used as data standardization to create alert format as standard as possible.

Basant Agarwal and Namita Mittal [13] proposed in their paper ―Hybrid Approach for Detection of Anomaly Network Traffic applying Data Mining methods a hybrid method that exploits the benefits of both the methods i.e. entropy based and support vector machine based respectively. The hybrid anomaly detection system learns the network traffic behavior from the normalized various network features entropy values. Entropy based methods have the better representing advantage the network traffic properties and SVM is decent for classification. The normalized entropies are SVM model for learning sent the network behavior. This trained SVM model can network traffic classify in legitimate traffic or attack traffic.

Augustin Orfila, Javier Carbo and Arturo Ribagorda [14] proposed a system based on multiagents to improve the overall IDS effectiveness through an autonomous adaptation in their paper ―Autonomous decision on Intrusion Detection with trained BDI agents‖. The system is composed of several cooperative agents that play one of the following roles: sensor, evaluator or manager. Each sensor agent applies a specific detection algorithm to infer a prediction about the intrusive nature of the attack. The predictions are often binary in statement indicating the intrusive or non-intrusive behaviors that are sent to evaluator agents. The evaluator agents further combine them to produce a final conclusion which is sent to the manager agent. Evaluator agents apply two different criteria to conclude the nature of attack. The two criteria that are considered are: Threshold: The evaluator agent considers an event as an intrusion if the number of sensor agents that state the event as intrusive is greater than a prefixed threshold.

Tadeusz Pietraszek et al. [15] proposed two complementary approaches CLARAty and ALAC to be utilized together in a two-staged alert filtering and classification system in their paper ― machine learning and Data Mining-Towards reducing false positives in intrusion detection‖. The proposed system uses CLARAty in first-stage to periodically mine raw alerts and discover their root causes. Then it would either remove them or install alert filters. The output of CLARAty would then be forwarded to ALAC interacting with an operator. The major benefit of this approach is that it alert filters from CLARAty remove the most prevalent and uninteresting false positives, which effectively improves class distribution in favour of true positives in the alerts passed on to the second stage. ALAC receives fewer alerts to process and is an adaptive alert classifier based on feedback of an intrusion detection analyst and machine learning techniques.

According to Cheng Xiang, Png Chin Yong and Lim Swee Menzns' [16] paper on ― multiple-level hybrid design classifier for intrusion detection system applying decision trees and Bayesian clustering detection rate can be increased by implementing a new multiple-level intrusion hybrid classifier. A model with 4 stages of classification is used for the hybrid classifier. The first level of classification categorizes the test data into 3 categories (DOS, Probe, Others). U2R and R2L and the Normal connections are classified as ―Others in this stage. The second stage splits ―Others‖ into Attack and Normal categories, while the third stage separates the Attack class from Stage 2 into U2R and R2L. The fourth stage further classifies the attacks into more specific attack types. This classification is only effective for known attacks as it requires that particular type of belabored training data to be present.

Xiao-Bai Li [17] in his paper ―A system of scalable decision tree and its application in the pattern intrusion and recognition detection‖ proposed a novelalgorithm of decision tree, named SURPASS, that is extremely efficient in handling large data. It is based on an efficient gathering of sufficient statistics. The algorithm effectively solves the problem of mining big numeric information for classification when the size of data is beyond the capacity of the basic memory. The algorithm is based on univariate or multivariate splits. It is specialized in dealing with numeric information. The outcomes represent that the proposed algorithm creates decision trees with most high feature in classification accuracy terms.

LNID is proposed for detecting attacks on Telnet traffic by ChiMei Chen et al [18] in their paper ―An efficient network intrusion detection‖. According to their proposed work, normal traffic behavior is taken into consideration and anomaly score of a packet based on deviation from the normal behavior is computed. Instead of processing eachpackets of traffic, an effective filtering method is dudes to reduce the system workload. The filtering scheme consists of 2 packet filters: Tcpdump filter and LNID filter. The former, processes initial packet filtering with tcpdump tool, extracting TCP packets towards Telnet servers of internal local area network.
Jaehak Yu et al. [19] proposed, implemented and designed a system that traffic flooding attacks detects and executes classification by the attack kind and uses SNMP MIB (Simple Network Management Protocol) MIB (Management Information Base) based on the C4.5 algorithm in their paper ―An in-depth analysis on traffic flooding attacks detection and system using data mining techniques‖. The proposed system is composed of 3 modules: SNMP MIB generators (for online processing) module, MIB update detection and MIB data store, attack detection and classification module and for offline processing, C4.5 training and association rule mining module and lastly the system administrator as a management module.

In this paper ―An active learning based TCM-KNN algorithm for supervised network intrusion detection‖, by Yang Li and Li Guo [20] a network of novel supervised intrusion detection technique based on the TCM-KNN algorithm of machine learning and active learning based training data selection method is proposed. It can efficiently detect anomalies with the high detection rate, low false positives under circumstance of applying much fewer selected data as well as selected features for the training in the comparison with classical methods of supervised intrusion detection. A progression of trial results on the surely understood KDD Cup 1999 information set show that the proposed strategy is more powerful and successful than the best in class interruption recognition routines.

In this paper ―Data-mining based SQL attack detection using internal query trees ‖, Mi-Yeon Kim and Dong Hoon Lee [21], proposed a system to recognize SQLIAs at database level by utilizing SVM grouping and different part works. Identifying SQL infusion assaults (SQLIAs) is turning out to be progressively vital in database-driven sites. A large portion of the studies on SQLIA identification have concentrated on the organized question dialect (SQL) structure at the application level and this methodology unavoidably neglects to recognize those assaults that utilization as of now put away strategy and information inside of the database framework. The prime issue of SQLIA discovery structure is the manner by which to speak to the interior inquiry tree gathered from database log suitable for SVM grouping calculation keeping in mind the end goal to gain great execution in identifying SQLIAs. To understand this issue, a novel strategy to change over the inquiry tree into a n-dimensional element vector by utilizing a multi-dimensional grouping as a moderate representation is proposed

In this study ―Application of ANN and SVM for intrusion detection‖, through Wun-Hwa Chen, Sheng-Hsun Hsu, HwangPin Shen [22] the feasibility of using an ANN and SVM to predict attacks based on the frequency-based encoding methods are determined. The aim of applying SVM and ANN for attack detection is to develop a generalization capability from limited training data. In addition to comparing the SVM and ANN performances, they demonstrated other encoding methods in predicting attacks. The test bed used here is 1998 DARPA data from MIT‘s Lincoln Labs. Results indicated that SVM performance was superior to that of ANN and the encoding technique is better than the simple frequency-based technique. The superior performance of SVMs over ANNs is due to the following reasons: (1) SVMs implement the structural risk minimization principle which minimizes an upper bound for the generalization error rather than minimizing the training error.

Levent Koc et al. [23] presented an intrusion detection model based on a multinomial classifier that is used to classify network events as normal or attack events, such as DoS, probe, U2R, and R2L. The model is based on a new data mining method called Hidden Naive Bayes (HNB). The HNB classifier model is applied to several datasets and shows promising results compared with the traditional Naive Bayes and its extended methods (Jiang, Zhang, & Cai, 2009). The experimental research study explores the traditional Naive Bayes and leading structurally extended Naive Bayes approaches including the new HNB approach. In the study, they augmented the Naive Bayes and extended Naive Bayes methods with the leading discretization and feature selection methods to increase the accuracy and decrease the resource requirements of intrusion detection problem. Based on the results of the proposed study, HNB based classifier model stands out as simple and practical intrusion detection system with better predictive accuracy and cost.

This study proposed by Ming-Yang Sua et al. [24] is a real-time NIDS with the incremental mining for the fuzzy association rules. While accepting fuzzy association rules to analyze traffic of network, particularly for a NIDS, the time expense, containg online information mining and collection techniques, are vital. Incremental fuzzy-rule mining is appropriate to meet real-time demands because it can create the latest rule set, while a novel knowledge record is gathered by internet. This paper first proposes an online incremental mining algorithm for creating fuzzy association rules, and then presents a real-time intrusion detection system based on the algorithm. By consistently comparing the two different rule sets, one mined from online packets and the other different mined from training attack-free packets, the proposed system can render a decision all 2 seconds. Thus, compared with classical static mining methods, the proposed system can greatly increase efficiency.

The proposed method through Tansel Ozyer, Reda Alhajj and Ken Barker [25] is based on iterative rule learning applying a fuzzy rule-based genetic classifier. The method is mostly composed of two different stages: first, a big number of candidate rules are generated for all class applying fuzzy association rules mining, and they are pre-screened applying two different rule evaluation criteria in order to decrease the search space of fuzzy rule. Candidate rules find after pre-screening are used in the genetic fuzzy classifier to create rules for the classes specified in the IIDS: PRB-probe, namely Normal, and U2R-user to root, DOS-denial of service and R2L-remote to local.

This study by Yogita B. Bhavsar et al. [26], proposes IDSusing data mining technique: SVM. Here, Classification is done through applying SVM and verification regarding the proposed system effectiveness will be done through conducting some experiments applying NSL-KDD Cup'99 dataset which is improved version of KDD Cup'99 data set. The SVM is one of the most prominent classification algorithms in the area of data mining, but its drawback is its extensive training time. In this proposed system, some experiments with applying NSL-KDD Cup'99 data set has been carried out.

The paper by Mrutyunjaya Pandaa, Ajith Abrahamb, and Manas Ranjan Patrac [27] uses a hybrid intelligent method applying classifiers combination in order to create the decision intelligently, so that the complete resultant model performance is enhanced. The commonprocess in this is to follow the un-supervised or supervised information filtering with classifier or cluster, first on the complete training data set and then the output is applied to another classifier to classify the data.

The purpose of this study by Dr. Saurabh Mukherjee, Neelam Sharma [28] is to identify significant reduced input features in building IDS that is computationally effective and efficient and for this the paper investigates the presentation of three basic feature selection approachesCFS, IG and GR. In this paper they proposed method Feature Vitality Based Reduction Technique, to identify significant summary input features. Then they applied one of the effiective classifier Naive Bayes on reduced information sets for intrusion detection. Experimental outcome illustrates feature subset identified throughCFS has improved Naive Bayes classification correctness when compared to GR and IG. Although GR is an extended of IG, but during analysis they have used both the methods for feature selection and IG performs higher than GR. FVBRM method present classification accuracy improve with compared to the CFS but takes more time. Empirical outcomes present that selected reduced attributes give higher performance to design IDS that is effective and efficient for network IDS.

In 2014, Mradul Dhakar and Akhilesh Tiwari [34] proposed a new hybrid model for IDS. It is a detection mechanism for detecting the intrusive activities unseen among the normal activities. They proposed a framework which may be expected as another different level towards IDS advancement. The framework use the crucial data mining classification algorithms useful for intrusion detection. It is framework of hybrid intrusion detection based on the combination of two different classifiers i.e. TAN and REP. The TAN classifier is used as a base classifier while the REP classifier is used as a Meta classifier. The development framework is an intelligent, adaptive and efficient intrusion detection framework.

VI. CONCLUSION

Since the ready-made data mining algorithms is presented, intrusion detection based on the data mining has developed rapidly. It advances in the ability to handle massive data, but it also has problems like, for instance,searching for more effective data mining algorithms, how to improve the correct rate of intrusion detection, how to control the rate of false alarm in anomaly detection and etc.These can be the topics for future research, meanwhile they also need lots of work and experiments to develop a system that is more effective and more appropriate. There are many types of approaches in intrusion detection, in which thatbased on the data mining becomes the hot spot in the present intrusion detection methodology. However, data mining is stillin its developing stage, so more thorough study needs to bedone. A brief survey of the IDS in the data mining field is given in this paper.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining: Concepts andTechniques”, 3rd edition, Morgan Kaufmann, 2011. (1st ed., 2000-2001) (2nd ed., 2006)

[2] B. Babcock, M. Datar, and R. Motwani ,”Load Shedding Techniquesfor Data Stream Systems” (shortpaper), Proc. of the 2003 Workshopon Management and Processing of Data Streams, June 2003I.S.Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. NewYork: Academic, 1963, pp. 271-350.

[3] Bifet, Albert. "Mining Big Data in Real Time", Informatica37,pp:15-20, 2013.

[4] B. Babcock, S. Babu, M. Datar, R. Motwani, and J.Widom,” Modelsand issues in data stream systems”, Proceedings of PODS, 2002.

[5] Zhen-Ya Zhang, Hong-Mei Cheng, et al. From data mining toOpportunity / symptoms found. Computer Science. 2007.

[6] Dai Yingxia, lian Yi-feng, Wang hang. Security and intrusiondetection systems [M]. Beijing: Tsinghua University Press, 2002.

[7] Manish Kumar, Dr. M. Hanumanthapaa, “Intrusion Detection Systemusing Stream Data Mining and Drift Detection Method”,4th lCCCNT -2013 July 4-6,2013, Tiruchengode,India.

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

[9] Nadianmai G. V., Hemalathain M., ―Effective approach toward Intrusion Detection System using data mining techniques‖, Cairo University, Elsevier, Egyptian Informatics Journal, 2014, pp. 37-50.

[10] Feng Wenying, Zhang Qinglei, Hu Gomgzhu, Huang Jimmy Xiangi, ―Mining network data for intrusion detection through combining SVMs with Ant Colony Networks‖, Elsevier, Future Generation Computer Systems 37(2014), pp. 127-140.

[11] Mohammad Muamer N., Sulaiman Norrozila , Muhsin Osama Abdulkarim, ―A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment‖, Elsevier, Procedia Computer Science 3(2011), pp. 1237-1242.

[12] Al-Saedi Karim, r ManickamSelvakuma, Ramadass Sureswaran, Al-Salihy Wafaa and ALmomani Ammar, ―Research Proposal: An Intrusion Detection system Alert Reduction and assessment Framework Based on Data Mining‖, Journal of Computer Science, 2013, ISSN 1549-3636, 9(4): 421-426.

[13] AgarwalBasan, Mittal Namita, ―Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques‖, Elsevier, Procedia Technology 6(2012)996-1003.

[14] Orfila Augustin, Carbo Javier, Ribagorda Arturo, ―Autonomous decision on Intrusion Detection with trained BDI agents‖, Elsevier, Computer Communications 31(2008): 1803-1813.

[15] Pietraszek Tadeusz, Tanner Axel, ―Data Mining and machine learning-Towards reducing false positives in intrusion detection‖, Elsevier, Information Security Technical Report (2005) 10:169-183.

[16] Xiang Cheng, Yong Png Chin Menz, Lim Swee, ―Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees‖, Elsevier, Pattern Recognition Letters 29 (2008) 918–924.

[17] Li Xiao-Bai, ―A scalable decision tree system and its application in pattern recognition and intrusion detection‖, Elsevier, Decision Support System 41(2005) 112-130.

[18] Chen Chia-Mei, Chen Ya-Lin Lin, Hsao-Chung, ―An efficient network intrusion detection‖, Elsevier, Computer Communications 33(2010) 477- 484.

[19] Yu Jaehak, Kang Hyunjoong, Park DaeHeon, Bang Hyo-Chan, Kang Do Wook, ―An in-depth analysis on traffic flooding attacks detection and system using data mining techniques‖, Elsevier, Journal of Systems Architecture 59(2013) 1005-1012.

[20] Li Yang, Guo Li, ―An active learning based TCM-KNN algorithm for supervised network intrusion detection‖, Elsevier, Computers & Security 26(2007) 459–467.

[21] Kim Mi-Yeon Lee, Dong Hoon, ―Data-mining based SQL injection attack detection using internal query trees‖, Elsevier, Expert Systems with Applications 41 (2014) 5416–5430.

[22] Chen Wun-Hwa, Hsu Sheng-Hsun, Shen Hwang-Pin, ―Application of SVM and ANN for intrusion detection‖, Elsevier, Computers & Operations, Research 32 (2005) 2617–2634.

[23] Koc Levent, Mazzuchi Thomas A., Sarkani Shahram, ―A network intrusion detection system based on a Hidden Naive Bayes multiclass classifier‖, Elsevier, Expert Systems with Applications 39 (2012) 13492–13500.

[24] Su Ming-Yang, Yu Gwo-Jong, Lin Chun-Yuen, ―A real-time network intrusion detection system for large-scale attacks based on an incremental mining approach‖, Elsevier, Computers & Security 28 (2009)301–309.

[25] Ozyer Tansel, Alhajj Reda, Barker Ken, ―Intrusion detection by integrating boosting genetic fuzzy classifier and data mining criteria for rule prescreening‖, Journal of Network and Computer Applications 30 (2007) 99–113.

[26] Bhavsar Yogita B.Waghmare Kalyani C, ―Intrusion Detection System Using Data Mining Technique: Support Vector Machine‖, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Vol. 3, Issue 3, March 2013.

[27] Pandaa, Ajith Abrahamb, Patrac Manas Ranjan, ―A Hybrid Intelligent Approach for Network Intrusion Detection‖, Mrutyunjaya International Conference on Communication Technology and System Design 2012, Procedia Engineering 30(2012) 1–9.

[28] Mukherjee Dr. Saurabh, Sharma Neelam, ―Intrusion Detection using Naive Bayes Classifier with Feature Reduction‖, C3IT-2012, Procedia Technology 4(2012)119 – 128.

[29] Jiawei Han, Micheline Kambar, Jian Pei, "Data Mining Concepts and Techniques" Elsevier Second Edition.

[30] Margaret H.Dunham, "Data Mining Introductory and Advanced Topics" Pearson Education.

[31] http://docs.oracle.com

[32] Pedro G. Espejo, Sebastian Ventura, and Francisco Herrera, "A Survey on the Application of Genetic Programming to Classification", Page: 121-144 VOL. 40, NO. 2 MARCH 2010.

[33] www.wikipedia.org

[34] Dhakar Mradul, and Akhilesh Tiwari. (2014)"A Novel Data Mining based Hybrid Intrusion Detection Framework." *Journal of Information and Computing Science*9.1: 037-048.