ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Some Studies in Intrusion Detection using Data Mining Techniques

Inadyuti Dutt [1], Dr. Samarjeet Borah [2]

Assistant Professor, Dept. of Computer Applications, B.P. Poddar Institute of Management & Technology, Affiliated to West Bengal University of Technology, West Bengal, India[1]

Associate. Professor, Dept. of Computer Application, Sikkim Manipal Institute of Technology, Affiliated to Sikkim Manipal University, Sikkim, India[2]

ABSTRACT: This paper surveys the present data mining techniques used in detecting intrusions in a computer networks. The basic objective of this survey paper is to review the current methodologies of data mining technique being used in Intrusion Detection System. The paper focuses on the usage of various methodologies of data mining technique like clustering, classification and other data mining rules. The results suggest that classification methodology is being widely used for solving intruder-based problems and Support Vector Machine (SVM) remains popular within this arena, for the researchers. Similarly, in clustering technique, statistical-based, conditional probability i.e. Bayesian clustering, and its native are used for categorizing attack from a non-attack. Even if these methodologies score well in intrusion detection, the hybrid models introduced generate good performances in lowering false alarms.

KEYWORDS: Intrusion Detection System (IDS), Data Mining, Support Vector Machine (SVM), Clustering, Classification, Decision tree, C4.5

I. INTRODUCTION

The usage of computer and its applications have undeniably increased in the past few decades. The increased reliance of government, military and commercial organizations on computer and its applications have equally increased the threats on the computing systems. As a result security of our computer systems and data is at continual risk. Due to the extensive growth of the Internet and increasing availability of tools and tricks for intruding and attacking networks, new challenges arise in order to combat external attacks. Such attacks external to these bodies are deliberate in action against data, software or hardware and can destroy, degrade, disrupt or deny access to a network computer system. Intrusions are such deliberate attacks.

In an attempt to guard against the unknown intrusions, much effort has been given in researching and developing Intrusion Detection System (IDS), which tries to filter out such attacks from the network traffic. IDS are software tools meant specifically for strengthening the security of information and communication systems. An IDS dynamically monitors logs and network traffic, applies detection algorithms to identify intrusions in a network. Intrusion detection is based on the assumption that intrusive activities are noticeably different from normal system activities and thus are identifiable. Intrusion detection is not used to replace prevention-based techniques such as authentication and access control; instead, it is intended to complement existing security measures. Intrusion detection system is therefore considered as a second line of defense for computer network systems to detect actions that bypass the security monitoring and control component of the system. Intrusion Detection Systems monitors and analyzes the events occurring in a computer and/or network system in order to detect signs of security problems and raise alarms. Basic approaches used are known pattern templates, threatening behavior templates, traffic analysis, statistical-anomaly detection and state-based detection.

Data Mining in Intrusion Detection

Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

patterns in data. In recent years, data mining techniques employed in intrusion detection have proved to be successful. Data mining is the search for valuable information within large volumes of data by systematically exploring underlying patterns, trends, and relationships hidden in available data.

Data mining is commonly employed within the area of intrusion detection to find the hidden patterns of intrusions and their relationship among each other. Data mining techniques for intrusion detection are namely: Clustering frequent pattern mining, classification, mining data streams etc. Data mining can help to learn from traffic data using supervised learning approach by learning precise models from past intrusions or unsupervised learning approach by identifying suspicious activities.

II. LITERATURE REVIEW

G. V. Nadianmai and M. Hemalathain in their paper "Effective approach toward Intrusion Detection System using data mining techniques", considered four issues namely Classification of Data, High Level of Human Interaction, Lack of Labeled Data, and Effectiveness of Distributed Denial of Service attack and solved them using the proposed algorithms EDADT algorithm, Hybrid IDS model, Semi-Supervised Approach and Varying HOPE RAA algorithm respectively. To solve the problem related to classification of data, an enhanced data adapted decision tree algorithm is implemented which effectively classifies the data into normal and attack without any classification. To minimize the workload of a network administrator, a high level of human interaction based on SNORT and anomaly based approaches are being used. This has a Hybrid IDS that automatically classifies the data based on the pre-defined rules within it. The issue related to belling the unlabeled data is solved using Semi-Supervised Approach where with the small amount of labeled data, the large amount of unlabeled data can be labeled. The last problem related to Distributed Denial of Service Attack is addressed by using 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 by the legitimate client.

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

Muamer N. Mohammad et al. use Weka software, which is a collection of machine learning algorithm for data mining tasks in their paper "A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment". Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. Weka consists of Explorer, Experimenter, Knowledge flow, Simple Command Line Interface. The proposed system has four main steps to execute. The first step is to collect network data and pre-treat them as network connection data including particular attributes as protocol type, destination IP address and flag bit. Then the next step is to use association analysis data mining algorithm to handle the connection data and get association rules, thereby obtaining the normal behavior patterns which can be used for abnormal intrusion detection. Finally, the proposed work uses classification algorithm to carry out rule mining to further distinguish normal behavior and intrusion behavior.

The proposed work by **Karim Al-Saedi et al.** in their paper "Research Proposal: An Intrusion Detection system Alert Reduction and assessment Framework Based on Data Mining" is IDS Alert Reduction and Assessment based on Data Mining (ARADMF) which contains three systems: Traffic data retrieval and collection mechanism system, reduction IDS alert processes system and threat score process of IDS alert system. The traffic data retrieval and collection mechanism systems develops a mechanism to save IDS alerts, extract the standard features as intrusion detection exchange format and save them in DB file(CSV-type). It contains the Intrusion Detection Message Exchange Format(IDMEF) which works as procurement alerts and field reduction is used as data standardization to make format of alert as standard as possible [4].

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

This algorithm consists of three phases. The first phase removes redundant alerts; the second phase reduces false alerts based on threshold time value and the last phase reduces false alerts based on rules with a threshold common vulnerabilities and exposure value. Threat score process of IDS alert system is characterized by assigning scores based on automated classification of alerts. The expected outcome reduces the number of false positive alert with rate expected 90% and increasing the level of accuracy compared with other approaches.

Basant Agarwal and Namita Mittal proposed in their paper "Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques" a hybrid approach that exploits the benefits of both the techniques i.e. entropy based and support vector machine based respectively. Hybrid anomaly detection system learns the behavior of network traffic from the normalized entropy values of different network features. Entropy based techniques have the advantage of better representing the properties of the network traffic and support vector machine is good for classification.

The normalized entropies are sent to SVM model for learning the behavior of the network. This trained SVM model can classify the network traffic in attack traffic or legitimate traffic. In entropy based anomaly detection system, firstly normalized entropy of network traffic features is calculated in every 60 seconds. Threshold value is fixed for each feature for identifying the anomalies based on experiments. Then voting system for each feature decides whether there is an attack or not. This method is able to produce good results in case of detecting attack traffic but it also produces high false alarms, because the entropy values can also deviate from the range or towards 0 or 1 in case of legitimate traffic [5].

Augustin Orfila, Javier Carbo and Arturo Ribagorda 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 [6].Weighted sum: the evaluator agent weighs each sensor before the comparison with the predefined threshold is done. The weights are updated after an event takes place according to the historical effectiveness of each sensor [6].

The manager agent finally may act in two different modes of behavior: evaluation mode or operating mode. In the former one it is assumed that the manager agent has the knowledge about the real nature of events, so it is able to inform the elevator agents about their effectiveness. The evaluator agents in turn will utilize this information in order to update the weights of sensor agents, for future predictions. The operating mode consists of planning a response to intrusions, according to the beliefs previously acquired about the environment where events are taking place and about the results provided in the evaluation mode.

Tadeusz Pietraszek et al. proposed two complementary approaches CLARAty and ALAC to be utilized together in a two-staged alert filtering and classification system in their paper "Data Mining and machine learning-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 [7]. ALAC receives fewer alerts to process and is an adaptive alert classifier based on feedback of an intrusion detection analyst and machine learning techniques.

Experiments with real-world data sets have shown that already few dozens of generalized alerts cover over 90% of the raw alerts [7].

According to **Cheng Xiang, Png Chin Yong and Lim Swee Menzns'** paper on "Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees" detection rate can be increased by

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

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 [8].

Xiao-Bai Li in his paper "A scalable decision tree system and its application in pattern recognition and intrusion detection" proposed a new decision tree algorithm, named SURPASS (for Scaling Up Recursive Partitioning with Sufficient Statistics), that is highly efficient in handling large data. It is based on an efficient gathering of sufficient statistics. The algorithm effectively solves the problem of mining large numeric data for classification when the data size is beyond the capacity of the main memory [9]. The algorithm is based on uni-variate or multivariate splits. It is specialized in dealing with numeric data. When categorical data are presented, categorical values need to be processed with binary (0-1) coding. When there exists many categorical attributes each having a large number of categories, the coding process involves creating a large number of additional binary attributes. This could cause computational problems and thus a more natural approach to deal with mixed data type is to handle numeric and categorical data using different sufficient statistics. For numeric data, the summation statistics could be used. The quality of split can be evaluated using the same impurity measure such as entropy, no matter whether the split is based on numeric or categorical attributes [9]. The results indicate that the proposed algorithm produces decision trees with very high quality in terms of classification accuracy and the algorithm scores well against large data sets, with computing time approximately linear in terms of magnitude of number of records in the data.

A Lightweight Network Intrusion Detection System (LNID) is proposed for detecting attacks on Telnet traffic by **Chi-Mei Chen et al** 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 all traffic packets, an efficient filtering scheme 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 [10]. The module "LNID Filter" fetches the first packets of each Telnet connection. A TCP connection is established after a 3-way handshaking which includes SYN, SYN-ACK and ACK. The module "Anomaly Scoring" computes the anomaly score based on each attribute characterizing the normal behavior. As attack packets are sent right after a connection is built, thus, the proposed method adopts first 48 byte packet data for anomaly behavior detection [10]. The proposed method uses the module "Post Process" to remove multiple alerts. If multiple anomaly packets arrive, then multiple-alerts have to be propagated. The results show that LNID has a simple and efficient anomaly score function for detecting 86.4% of U2R and R2L attacks in Telnet connections. LNID is restricted to TELNET and has highest detection rate on R2L and U2R.

Jaehak Yu et al. proposed, designed and implemented a system that detects traffic flooding attacks and executes classification by the attack type and uses SNMP MIB (Simple Network Management Protocol) MIB (Management Information Base) based on 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. SNMP MIB generator's module generates MIB information from the network traffic data;MIB update detection module stores only the MIB information that is determined in the C4.% training module from the target system.; the collected information is transferred to attack detection, classification module and then it is used to judge the occurrence of attacks and the attack type in real-time; C4.% training module randomly generates various traffic attacks to execute a C4.5-based learning; Association rule module conducts an in-depth semantic interpretation that extracts and analyzes the data characteristics

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

of the data stored in the MIB data store module in a form of rule and System management module detects traffic flooding attack in real-time. Monitors detailed information about classification type [11].

In this paper "An active learning based TCM-KNN algorithm for supervised network intrusion detection", by **Yang Li and Li Guo** a novel supervised network intrusion detection method based on TCM-KNN (Transductive Confidence Machines for K-Nearest Neighbors) machine learning algorithm and active learning based training data selection method is proposed. It can effectively detect anomalies with high detection rate, low false positives under the circumstance of using much fewer selected data as well as selected features for training in comparison with the traditional supervised intrusion detection methods [12]. A series of experimental results on the well-known KDD Cup 1999 data set demonstrate that the proposed method is more robust and effective than the state-of-the-art intrusion detection methods.

TCM-KNN algorithm is commonly used machine learning and data mining method, thus effective in fraud detection, pattern recognition and outlier detection. It is the first time that TCM-KNN algorithm is applied to intrusion detection. Contrast experimental results demonstrate that it has good detection performance (high detection rate and low false positives) even when provided with ''small'' data set for training than the state-of-the-art intrusion detection technique [12]. An active learning based TCM-KNN algorithm for supervised network intrusion [12]. They further optimize it for intrusion detection in two aspects: (a) introduce active learning method to select much fewer good quality data for training than traditional random sampling, thus alleviate the large amounts of labeling workload for domain experts and reduce the scale of training data set, and consequently reduce the computational cost of TCM-KNN, and (b) feature selection method is proposed to select the most necessary and important features for TCM-KNN.

In this paper "Data-mining based SQL injection attack detection using internal query trees", **Mi-Yeon Kim and Dong Hoon Lee**, proposed a framework to detect SQLIAs at database level by using SVM classification and various kernel functions. Detecting SQL injection attacks (SQLIAs) is becoming increasingly important in database-driven websites. Most of the studies on SQLIA detection have focused on the structured query language (SQL) structure at the application level and this approach inevitably fails to detect those attacks that use already stored procedure and data within the database system. The prime issue of SQLIA detection framework is how to represent the internal query tree collected from database log suitable for SVM classification algorithm in order to acquire good performance in detecting SQLIAs. To solve this issue, a novel method to convert the query tree into an n-dimensional feature vector by using a multi-dimensional sequence as an intermediate representation is proposed. The reason that it is difficult to directly convert the query tree into an n-dimensional feature vector is the complexity and variability of the query tree structure. Secondly a method to extract the syntactic features, as well as the semantic features when generating feature vector is proposed. Next they proposed a method to transform string feature values into numeric feature values, combining multiple statistical models. The combined model maps one string value to one numeric value by containing the multiple characteristic of each string value [13]. In order to demonstrate the feasibility of the proposals in practical environments, they implemented the SQLIA detection system based on PostgreSQL, a popular open source database system, and the experimental results using the internal query trees of PostgreSQL validate that the proposal is effective in detecting SQLIAs, with at least 99.6% of the probability that the probability for malicious queries to be correctly predicted as SQLIA is greater than the probability for normal queries to be incorrectly predicted as SQLIA.

In this study "Application of SVM and ANN for intrusion detection", by **Wun-Hwa Chen, Sheng-Hsun Hsu, Hwang-Pin Shen** the feasibility of applying an Artificial Neural Network (ANN) and Support Vector Machine (SVM) to predict attacks based on frequency-based encoding techniques are determined. The goal of using ANN and SVM for attack detection is to develop a generalization capability from limited training data. In addition to comparing the ANN and SVM 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 method is better than the simple frequency-based method. 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.

However, ANNs implement the empirical risk minimization principle, which might lead to worse generalization than

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

SVM. (2) An ANN may not converge to global solutions due to its inherent algorithm design. In contrast, solutions in SVMs are equivalent to solving a linearly constrained quadratic programming problem, which leads to a global optimal solution. (3) In choosing parameters, SVMs are less complex than ANNs [15].

In this paper by **Gisung Kim et al.**, the C4.5 decision tree (DT) is used to create the misuse detection model and the one-class support vector machine (1-class SVM) is used to create multiple anomaly detection models. In order to implement the concept described above, the DT model is first trained based on a training data set consisting of normal traffic and known attack traffic information, and then a 1-class SVM model is trained for each normal training data subset decomposed by the DT model [15]. The proposed hybrid intrusion detection method was evaluated by conducting experiments using the NSLKDD data set, which is a modified version of the famous KDD Cup 99 data set and the experiment results demonstrate that the proposed method is better in terms of detection rates for unknown and known attacks than the conventional methods that independently train the DT model and 1-class SVM model without the proposed decomposition technique.

Levent Koc et al. 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.** is a real-time NIDS with incremental mining for fuzzy association rules. While adopting fuzzy association rules to analyze network traffic, especially for a NIDS, the time expense, including online data collection and mining procedures, are vital. Incremental fuzzy-rule mining is suitable to meet real-time demands because it can produce the latest rule set, while a new data record is gathered online. This paper first proposes an online incremental mining algorithm for generating fuzzy association rules, and then presents a real-time intrusion detection system based on the algorithm. By consistently comparing the two rule sets, one mined from online packets and the other mined from training attack-free packets, the proposed system can render a decision every 2 seconds. Thus, compared with traditional static mining approaches, the proposed system can greatly improve efficiency from offline detection to real-time online detection. As the proposed system derives features from packet headers only, like the previous works based on fuzzy association rules, large-scale attack types are focused. Many DoS attacks were experimented in this study. Experiments were performed to demonstrate the excellent effectiveness and efficiency of the proposed system.

The proposed method by **Tansel Ozyer, Reda Alhajj and Ken Barker** is based on iterative rule learning using a fuzzy rule-based genetic classifier. The approach is mainly composed of two phases: first, a large number of candidate rules are generated for each class using fuzzy association rules mining, and they are pre-screened using two rule evaluation criteria in order to reduce the fuzzy rule search space. Candidate rules obtained after pre-screening are used in genetic fuzzy classifier to generate rules for the classes specified in IIDS: namely Normal, PRB-probe, and DOS-denial of service, U2R-user to root and R2L-remote to local. During the next stage, boosting genetic algorithm is employed for each class to find its fuzzy rules required to classify data each time a fuzzy rule is extracted and included in the system. Boosting mechanism evaluates the weight of each data item to help the rule extraction mechanism focus more on data having relatively more weight, i.e., uncovered less by the rules extracted until the current iteration.

This study by **Yogita B. Bhavsar et al.**, proposes Intrusion Detection System using data mining technique: SVM (Support Vector Machine). Here, Classification is done by using SVM and verification regarding the effectiveness of the proposed system will be done by conducting some experiments using 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 data mining area, but its drawback is its extensive training time. In this proposed system, some experiments using NSL-KDD

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Cup'99 data set has been carried out. The experimental results show that it can reduce extensive time required to build SVM model by performing proper data set pre-processing. Also when proper selection of SVM kernel function is done such as Gaussian Radial Basis Function, attack detection rate of SVM is increased and False Positive Rate (FPR) is decreased.

The paper by **Mrutyunjaya Pandaa, Ajith Abrahamb, and Manas Ranjan Patrac** uses a hybrid intelligent approach using combination of classifiers in order to make the decision intelligently, so that the overall performance of the resultant model is enhanced. The general procedure in this is to follow the supervised or un-supervised data filtering with classifier or cluster, first on the whole training data set and then the output is applied to another classifier to classify the data [20]. The 2-class classification strategy along with 10-fold cross validation method to produce the final classification results in terms of normal or intrusion. Experimental results on NSL-KDD data set, an improved version of KDDCup 1999 dataset show that the proposed approach is efficient with high detection rate and low false alarm rate. The advantage of applying such a methodology is that the combination strategy produces a similarity score or probability produced by each classifier, the distribution of which reveals how confident it is in making an intelligent decision. Secondly, the approach will learn from synthesized feature vectors of several classifiers and can make a better decision than any individual strategy.

The purpose of this study by **Dr. Saurabh Mukherjee, Neelam Sharma** is to identify important reduced input features in building IDS that is computationally efficient and effective and for this the paper investigates the performance of three standard feature selection methods using Correlation-based Feature Selection (CFS), Information Gain (IG) and Gain Ratio (GR). In this paper they proposed method Feature Vitality Based Reduction Method, to identify important reduced input features. Then they applied one of the efficient classifier Naive Bayes on reduced data sets for intrusion detection. Experimental result illustrates feature subset identified by CFS has improved Naive Bayes classification accuracy when compared to IG and GR. Although GR is an extended of IG, but during analysis they have used both the techniques for feature selection and IG performs better than GR. FVBRM method shows much more improvement on classification accuracy with compared to CFS but takes more time. Empirical results show that selected reduced attributes give better performance to design IDS that is efficient and effective for network intrusion detection.

Ahmed et al. suggested a combination of Data Mining (DM) and Network Behavior Analysis (NBA) to overcome the limitations in current IDS. NBA can help cover the gap in traditional network systems, which considers a good move for most of industries to integrate NBA with advanced DM to achieve a better performance. NBA can significantly enhance the value of the data generated from IDS that use DM as intrusion detection technique by analyzing and correlating large amount of sequence data. The system is composed of the following units:

• Computer network sensors: collect audit data and network traffic events and transmit these data to ID units.

• DM-ID unit: contains different modules that employ various DM algorithms and techniques (e.g., classification, clustering, etc.). Each module works independently to detect intrusions in the network traffic data [22]:

•NBA-ID unit: deploys NBA to detect intrusions in the network audit data.

• Intrusions collection unit: collects detected intrusions from DM-ID and NBA-ID units.

• Visualization unit: help monitor and visualize the results of ID units.

• Managerial decision maker: analyzes intrusion results, evaluates system performance, takes decisions on detected intrusions, checks for false positives and false negatives, controls system operation, generates a performance report and decides if any changes/updates are needed.

In this paper by **Lih-Chyau Wuu, Chi-Hsiang Hung, Sout-Fong Chen** the intrusion pattern discovery module arguments Snort v1.9 with the ability to automatically mine single and intrusion patterns from attack packets, and the Intrusion Behavior Detection Engine extends the detecting ability of Snort v1.9 for sequential intrusion behavior. Furthermore, the intrusion patterns generated by our miner not only is applied for Snort system but also can be easily

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

ported into other detecting systems. Administrator can adapt min _sup values for different attacking programs by our menu-driven interface to find optimal intrusion patterns.

The miners are based on Apriori algorithm. Such algorithm needs to scan database repeatedly to mine frequent patterns. .The contributions of this work area by **Panos Louvieris et al.**, is that it combines the techniques namely k-means clustering, Naive Bayes, Kruskal–Wallis and C4.5; allows cyber attacks, as anomalies to be pin pointed with a high degree of accuracy within the cluttered and conflicted cyber network environment. Further, the inclusion of the Naive Bayes feature selection and the Kruskal–Wall is test in the approach facilitates the classification of both statistically significant and relevant feature sets, including a statistical benchmark for the validity of the approach [24]. In addition to this, the efficiency of the analysis by speeding up and reducing the amount of processing required is increased. Moreover, the power of the intrusion detection approach developed here in is its applicability for detecting different types of cyber attack.

Srilatha Chebrolu, Ajith Abraham and Johnson P. Thomas have investigated new techniques for intrusion detection and performed data reduction and evaluated their performance on the DARPA benchmark intrusion data. The feature selection method using Markov blanket model and decision tree analysis have been used. Following this, they explored the general Bayesian network (BN) classifier and Classification and Regression Trees (CART) as intrusion detection models and have also demonstrated performance comparisons using different reduced data sets. The proposed ensemble of BN and CART combines the complementary features of the base classifiers. Finally, a hybrid architecture involving ensemble and base classifiers for intrusion detection has been proposed. From the empirical results, it is seen that by using the hybrid model Normal, Probe and DOS could be detected with 100% accuracy and U2R and R2L with 84% and 99.47% accuracies, respectively.

The proposed paper by **Yogita B. Bhavsar et al.** uses SVM method for classification, which can reduce the time required to build model for classification and increase the intrusion detection accuracy when Gaussian RBF kernel is used. The experimental results show that, when data sets are properly processed and proper SVM kernel is selected i.e. Radial Basis Function (RBF), it can overcome the drawback of SVM i.e. extensive time required to build model. They conducted experiment with 10 fold cross validation and Gaussian RBF kernel of SVM, the time required to build model was 77.07 seconds and attack detection accuracy achieved was 94.1857 %. This attack detection accuracy was increased to 98.5749 %, when they changed classification to 10 fold cross validation and re-evaluation using supplied test set with same RBF SVM kernel function.

III. DISCUSSIONS

Clustering techniques in [3], [20] and [22] use to classify the data according to some behavior or some hidden patterns. Clustering is usually used in grouping a population and where little is known about the spread or relationships within the data. Clustering analysis is one of the techniques where objects are grouped into natural groupings based upon the characteristics they possess and each item in a cluster is similar to those within its cluster, but is dissimilar to those in other clusters. It segregates the attacks by clustering similar attack instances together or by highlighting them as outliers to the cluster assignment. There are many different clustering methods, although partition-based clustering is often applied. Partition-based clustering methods partition the data points into a number of clusters. It iteratively rearranges data points in an initial grouping of a predefined number of clusters. Although the performance of the methods relies on the initial arrangement of cluster centroids and also repeated run scan overcome this problem. Popular examples include K-means clustering technique used in [24] which tries to partition a set of points into K sets or clusters such that points in each such cluster tend to near each other. It is unsupervised as the points have no prior classifications.

Another technique based upon fuzzy set theory in [17], [18] where reasoning is approximated rather than precisely deduced from predicate logic. Fuzzy clustering has been employed in intrusion detection by a number of researchers.

Bayesian clustering methods in [8] [21] focus on the computation of class-conditional probability density. They are commonly used in the statistics community. N**aive Bayes classifiers** used in [21], [24] are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes was introduced under a different name into the text retrieval community in the early 1960s, and

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. In the statistics and computer science literature, Naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method.

Bayes essentially ignores attribute dependencies and is often violated. On the other hand, although a Bayesian network can represent arbitrary attribute dependencies, learning an optimal Bayesian network from data is intractable. The main reason is that learning the optimal structure of a Bayesian network is extremely time consuming. Thus, a Bayesian model without structure learning is desirable. In this paper [16], a novel model, called *hidden naive Bayes* (HNB) has been proposed. In an HNB, a hidden parent is created for each attribute which combines the influences from all other attributes. HNB inherits the structural simplicity of naive Bayes and can be easily learned without structure learning.

Classification techniques used in [7], [20], [22] in order to identify a model that describes a data class or concept, categorizing data items into classes according to their pre-defined attributes. For example, in order to identify 'attack' or 'not attack', it is important to detect whether the data is specifically, 'trojan' or not. Classification is a supervised learning technique, which uses a training set with predefined labels with which to classify data items However, labeled intrusion data can be very difficult to obtain because audit data analysis is time consuming and attack instances occur irregularly.

Support Vector Machine (SVM) in [2], [5], [13], [14], [15], [19] and [26] has recently been introduced as a new technique for solving a variety of learning, classification and prediction problems. The basic SVM deals with two-class problems—in which the data are separated by a hyperplane defined by a number of support vectors. Support vectors are a subset of training data used to define the boundary between the two classes. In situations where SVM cannot separate two classes, it solves this problem by mapping input data into high-dimensional feature spaces using a kernel function. In high-dimensional space it is possible to create a hyperplane that allows linear separation (which corresponds to a curved surface in the lower-dimensional input space). Accordingly, the kernel function plays an important role in SVM. Decision tree in [1], [8], [9], [15] and [25] is one of the most well-known data mining techniques. Classification tree and Regression tree (CART) in [3], [7], [22] and [25] analyses two types of decision trees. In classification tree analysis, the predicted outcome is the class to which the data belongs whereas in regression tree analysis the predicted outcome can be considered a real number (e.g. the salary of an employee, or his experience in years). It is also widely used within intrusion detection as the decision rules created from the trees can then be used to update existing rule-based defences (e.g. Snort).

K-nearest neighbors in [12] is a classification (or regression) algorithm that in order to determine the classification of a point, combines the classification of the K nearest points. It is supervised because you are trying to classify a point based on the known classification of other points.

C4.5 used in [11], [15] [21] and [24] is an algorithm used to generate a decision tree C4. and it is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

Association rules in [3], [4] and [11] are another data mining technique which use if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to purchase milk. Apriori algorithm is a type of assocation rule and is the best-known algorithm to mine association rules. In [23], it uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support.

Biologically inspired neural network, agent-based and genetic-based algorithms are widely used in intrusion detection in recent times. They are model-based approaches which can greatly influence the overall performance of IDS.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Artificial Neural Network inspired by neural network in [14] is composed of simple processing units, or nodes, and the connection between any two units has some weight. A subset of the units acts as input nodes and another subset acts as output nodes, which perform summation and thresholding. In the neural network approach to intrusion detection, the neural network learns to predict the behavior of the various users and intruders in the system. The main advantage of neural networks is their tolerance to imprecise data and uncertain information. It has the innate ability to infer solutions from data without having prior knowledge of the regularities in the data.

Other approaches based on Ant Colony Optimization or Swarm Intelligence is used to employ agents in order to detect dynamic intrusion in real-time networks. Ant Colony [2], [6] based approaches incorporate agents that randomly move in a particular path to detect the intrusions in a networking environment.
Genetic-based approaches [18] have the ability to detect intrusions by using genetic operations like crossover, mutation and by determining fitness solutions for a particular optimization problem.

IV. RESULTS

The following figures depict the usage of Data mining techniques in IDS. From Fig. 1, it is evident that classification approach towards data mining is maximally used in detecting intrusion as because it has the innate ability to use the pre-defined labeled data for categorization and also lesser time consuming. Whilst the other techniques are being used in IDS, the usage of classification is more popular among researchers.

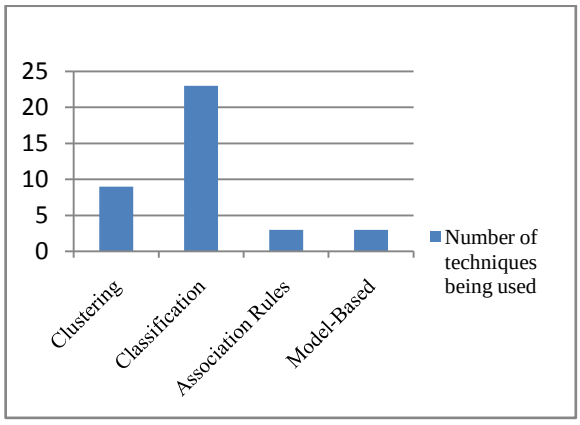

Fig. 1 Usage of various Data Mining Techniques in Intrusion System

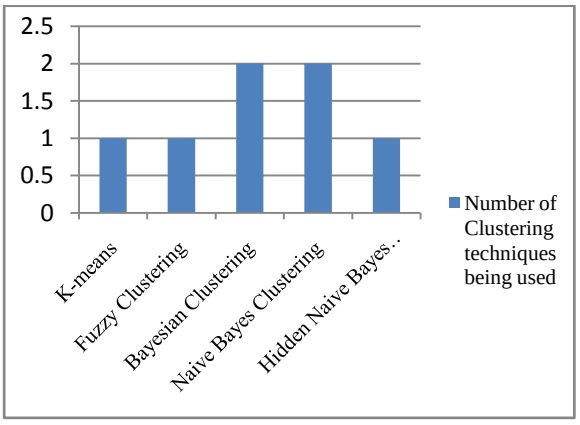

Fig. 2 Usage of various Clustering techniques in Intrusion Detection System

From Fig. 2, one could realize the usage of Bayesian, its native Naïve and Hidden Naïve Bayes clustering approaches being frequently used for grouping the unknown attack. They use conditional probability to group the known attacks from the unknowns. As K-means clustering and fuzzy methods consume more time in calculating the correct grouping thus Bayesian clustering is widely used.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

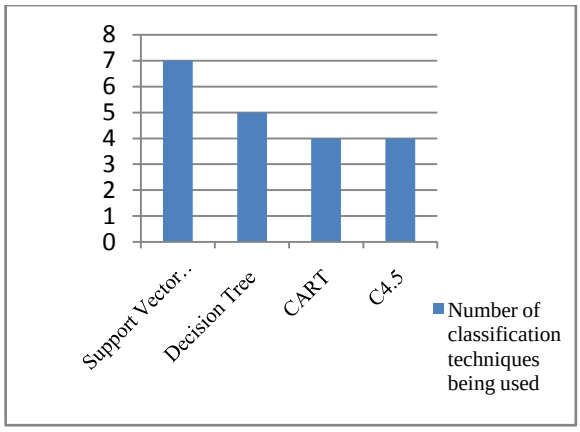

Fig. 3 Usage of Classification Techniques in Intrusion Detection System

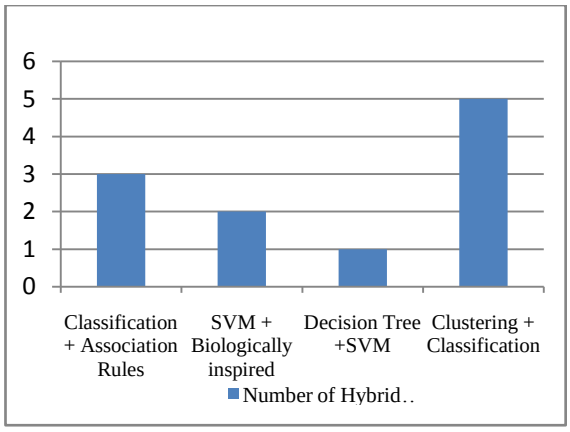

Fig. 4 Usage of Hybrid Models in Intrusion Detection System

In Fig. 3, the usage of various classification techniques are shown. It is seen that SVM is maximally being introduced for classification of data items in IDS. In situations where SVM cannot separate two classes, it solves this problem by mapping input data into high-dimensional feature spaces using a kernel function.
Even when each of these methods has been shown to perform well individually in specific scenarios, it is evident that their contributions increase when combined together in intrusion detection systems. Fig. 4 shows hybrid models that are usually based upon the combination of multiple clustering and classification techniques, with clustering performed first in order to remove the noise of the dataset. Various combinations of these techniques have been successfully employed for intrusion detection. For example, combined K-means clustering and ID3 decision trees is used to achieve very low false-positive rates. Further, combining Naïve Bayes and C4.5 decision tree classification reduces the false positive alarm rate.

V. CONCLUSIONS

The outcomes of this research paper suggest that there is an extensive usefulness and necessity of using Data Mining in Intrusion Detection System. The results also suggest that classification is widely in use for solving intruder-based problems and Support Vector Machine (SVM) remains popular with the researchers. Similarly, in clustering technique, statistical-based, conditional probability i.e. Bayesian clustering, and its native are used for categorizing attack from a non-attack. Lastly, it is also prevalent that various hybrid models are introduced in detecting intrusion. Hybrid models based on neural, agent or genetic algorithms are preferably used along with data mining techniques in order to detect intrusion in the system.

REFERENCES

[1] 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.
[2] 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.
[3] 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.
[4] 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.
[5] t AgarwalBasan, Mittal Namita, "Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques", Elsevier, Procedia Technology 6(2012)996-1003.
[6] Orfila Augustin, Carbo Javier, Ribagorda Arturo, "Autonomous decision on Intrusion Detection with trained BDI agents", Elsevier, Computer Communications 31(2008): 1803-1813.
[7] 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.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

[8] 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.
[9] 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.
[10] Chen Chia-Mei, Chen Ya-Lin Lin, Hsao-Chung, "An efficient network intrusion detection", Elsevier, Computer Communications 33(2010) 477-484.
[11] 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.
[12] Li Yang, Guo Li, "An active learning based TCM-KNN algorithm for supervised network intrusion detection", Elsevier, Computers & Security 26(2007) 459–467.
[13] 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.
[14 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.
[15] Kim Gisung, Lee Seungmin, Kim Sehun, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection", Elsevier, Expert Systems with Applications 41 (2014) 1690–1700.
[16] 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.
[17] 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.
[18] Ozyer Tansel, Alhajj Reda, Barker Ken, "Intrusion detection by integrating boosting genetic fuzzy classifier and data mining criteria for rule pre-screening", Journal of Network and Computer Applications 30 (2007) 99–113.
[19] 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.
[20] 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.
[21] Mukherjee Dr. Saurabh, Sharma Neelam, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction", C3IT-2012, Procedia Technology 4(2012)119 – 128.
[22] Youssef Ahmed, Emam Ahmed, "Network Intrusion Detection Using Data Mining and Network Behavior Analysis", International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011.
[23] Wuu Lih-Chyau, Hung Chi-Hsiang Chen, Sout-Fong, "Building intrusion pattern miner for Snort network intrusion detection system", The Journal of Systems and Software 80 (2007) 1699–1715.
[24 Louvieris] Panos, Clewley Natalie, Liu Xiaohui,"Effects-based feature identification for network intrusion detection", Neurocomputing, Neurocomputing 121(2013)265–273.
25] Chebrolu Srilatha, Abraham AjithThomas, Johnson P.,"Feature deduction and ensemble design of intrusion detection systems", Elsevier, Computers & Security (2005) 24, 295-307.
[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, Vol. 3, Issue 3, March 2013.