ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

Data Mining Techniques for Intrusion Detection: A Review

Abhaya[1], Kaushal Kumar[2], Ranjeeta Jha[3], Sumaiya Afroz[4]

M.Tech (IS), Department of Computer Scince & Engineering, Birla Institute of Technology, Mesra (Ranchi), India[1,3,4]

M.Tech (SE), Department of Software Engineering, Delhi Technological University, Delhi, India[2]

Abstract: With the dramatically development of internet, Security of network traffic is becoming a major issue of computer network system. Attacks on the network are increasing day-by-day. The most publicized attack on network traffic is considered as Intrusion. Intrusion detection system has been used for ascertaining intrusion and to preserve the security goals of information from attacks. Data mining techniques are used to monitor and analyze large amount of network data & classify these network data into anomalous and normal data. Since data comes from various sources, network traffic is large. Data mining techniques such as classification and clustering are applied to build Intrusion detection system. An effective Intrusion detection system requires high detection rate, low false alarm rate as well as high accuracy. This paper presents the review on IDS and different Data mining techniques applied on IDS for the effective detection of pattern for both malicious and normal activities in network, which helps to develop secure information system.

Keywords: Intrusion Detection System; Anomaly Detection; Misuse Detection; Data mining; Clustering; Classifications

I. INTRODUCTION

With the speedy escalation of Internet, there is enhancement in the lifestyle of people but at the cost of threats, which are created by either individuals or any organization. They are used to break the security of network.

Security means degree of protection given to the network or system. The main goals of security are confidentiality, Integrity and availability of data [1]. Attacks on network can be referred as Intrusion. Intrusion means any set of malicious activities that attempt to compromise the security goals of the information.

In early days, only conventional approaches were used for network such as encryption, firewalls, virtual private network etc but they were not enough to secure network completely. It is difficult to depend completely on static defense techniques. This increases the need for dynamic technique, which can be monitors system and identify illegal activities. Thus to enhance the network security dynamic approach is introduced and known as Intrusion Detection System. Intrusion detection system collects online information from the network after that monitors and analyzes these information and partitions it into normal & malicious activities, provide the result to system administrator [2].

IDS is the area, where Data mining is used extensively, this is due to limited scalability, adaptability and validity. In IDS data is collected from various sources like network log data, host data etc. Since the network traffic is large, the analysis of data is too hard. This give rise to the need of using IDS along with different Data mining techniques for intrusion detection . Lee & Salvatore J. Stolfo, Columbia University were first to apply Data mining techniques in the IDS [3]. Data mining techniques such as classification and clustering easily extract the information from large dataset.

The remaining part of the paper is structured in this way. Section I introduction, Section II review the related work on IDS using Data mining techniques, Section III explanation of IDS. In Section IV, Data mining and its techniques which are used in IDS are described and finally Section V brings us to the conclusion.

II. LITERATURE REVIEW

Xiang M.Y. Chang et.al.(2004) [4], designed a multiple-level tree classifier for Intrusion detection system and increase the detection rate. Classifier is more efficient in case of known attacks but for unknown vulnerabilities it gives low detection rate. **Peddabachigiri S. et.al.**(2007)[5], proposed a model of intrusion detection system combining decision tree and support vector machine (DTSVM) classification techniques and produces high detection rate. **Mrutyunjaya panda et. al.**(2008)[6], compares different data mining techniques for intrusion detection system and found that accuracy & performance of Naïve bayes classifier for all classes is better than the accuracy obtained in the case of different Decision tree algorithm but Decision tree is robust in detecting unknown intrusions in comparison to Naïve bayes classification algorithm. **M.Govindarajan et.al.**(2009)[7], proposed new K-nearest neighbour classifier applied on Intrusion detection system and evaluate performance in term of Run time and Error rate on normal and malicious dataset. This new classifier is more accurate than existing K-nearest neighbour classifier. **Mohammadreza Ektela et.al.**(2010)[8], used Support Vector Machine and classification tree Data mining technique for intrusion

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

detection in network. They compared C4.5 and Support Vector Machine by experimental result and found that C4.5 algorithm has better performance in term of detection rate and false alarm rate than SVM, but for U2R attack SVM performs better. **Song Naiping et.al.**(2010)[9], studied on Intrusion detection based on Data mining. Here, types of IDS means Misuse detection and Anomaly detection are described by the author along with different Data mining techniques which are used to build IDS. **T. Velmurugan et.al.**(2010)[10], compute the complexity between k-means and k-medoids clustering algorithm for uniform and normal distribution of data points and concluded that average time taken by k-Means algorithm is more in both the cases.

P. Amudha et.al.(2011)[11], observed that Random forest gives better detection rate, accuracy and false alarm rate for Probe and DOS attack & Naive Bayes Tree gives better performance in case of U2R and R2L attack. Also the execution time of Naive Bayes Tree is more as compared to other classifier. **Deepthy k Denatious et.al.**(2012)[1], describe different data mining techniques applied for detecting intrusions. Also describe the classification of Intrusion detection system and its working. For large amount of network traffics, clustering is more suitable than classification in the domain of intrusion detection because enormous amount of data needed to collect to use classification.

R. China Appala Naidu et.al.(2012)[12], used three Data mining techniques SVM, Ripper rule and C5.0 tree for Intrusion detection and also compared the efficiency. By experimental result, C5.0 decision tree is efficient than other. All the three Data mining technique gives higher than 96% detection rate. **Roshan Chitrakar et.al.**(2012)[13], proposed a hybrid approach to intrusion detection by using k-Medoids clustering with Naïve Bayes classification and observed that it gives better performance than K-Means clustering technique followed by Naïve Bayes classification but also time complexity increases when increase the number of data points.

Roshan Chitrakar et.al.(2012)[14], proposed a hybrid approach of combining k-Medoids clustering with Support Vector Machine classification technique and produced better performance compared to k-Medoids with Naïve Bayes classification. The approach shows improvement in both Accuracy and Detection Rate while reducing False Alarm Rate as compared to the k- Medoids clustering approach followed by Naïve bayes classification technique. **Sumaiya Thaseen et.al.**(2013)[15], analyzed different tree based classification techniques for IDS. Experimental results show that Random tree model reduces false alarm rate and has highest degree of accuracy.

III. INTRUSION DETECTION SYSTEM

The concept of IDS was proposed by Denning(1987), to identify, detect and trace the intrusion[11]. An IDS is a combination of software and hardware which are used for detecting intrusion[1]. It gathers and analyzes the network traffic & detect the malicious patterns and finally alert to the proper authority. The main function of IDS includes:[16]

- Monitoring and analyzing the information gathered from both user and system activities.
- Analyzing configurations of system and evaluating the file integrity and system integrity.
- For static records, it finds out the abnormal pattern.
- To recognize abnormal pattern, it use static records and alert to system administrator.

A. Classification of IDS

According to techniques used for intrusion detection based on whether attack's patterns are known or unknown, IDS classified into two category [2][17]:

(1) Misuse detection
(2) Anomaly detection

Misuse detection: It is Signature based IDS where detection of intrusion is based on the behaviors of known attacks like antivirus software. Antivirus software compares the data with known code of virus. In Misuse detection, pattern of known malicious activity is stored in the dataset and identify suspicious data by comparing new instances with the stored pattern of attacks.

Anomaly detection: [16][18][1] It is different from Misuse detection. Here baseline of normal data in network data in network eg load on network traffic, protocol and packet size etc is defined by system administrator and according to this baseline, Anomaly detector monitors new instances. The new instances are compared with the baseline, if there is any deviation from baseline, data is notified as intrusion. For this reason, it is also called behavior based Intrusion detection system.

TABLE 1
COMPARISON BETWEEN MISUSE DETECTION AND ANOMALY DETECTION

Signature – Based(Misuse Detection)	**Behaviour–Based(Anomaly Detection)**
Advantages	**Advantages**
-Higher Detection rate, Accuracy for known behaviors. -Simplest and effective method. -Low False alarm rate.	-can examine unknown and more complicated intrusions. - Rate of Missing report is low. -Detect new and unforeseen vulnerabilities.
Disadvantages	**Disadvantages**
- It can detect only known attacks. - Needs a regular update of the rules which are used. - Often no differentiation between an attack attempt and a successful attack. - Rate of Missing report is high.	- Needs to be trained and tuned model carefully, otherwise it tends to false – positives -low detection rate and high false alarm rate. - It can't identify new attacks because intrusion detection depends upon latest model.

B. Working of Intrusion Detection System [1]

Author presents 4-steps for working of IDS.

1) *Data Acquisition:* Data is collected from various sources by using particular software.
2) *Feature Selection:* Huge amount of data is collected from network traffic. So dataset for IDS becomes

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

large. For working on large dataset generate feature vectors, which contains only necessary data.

3) *Analysis:* In this step, Collected data is analyzed to determine whether data is suspicious or not. Here, various Data mining techniques are used for Intrusion detection.
4) *Action:* IDS alarms the administrator about attack which has been detected.

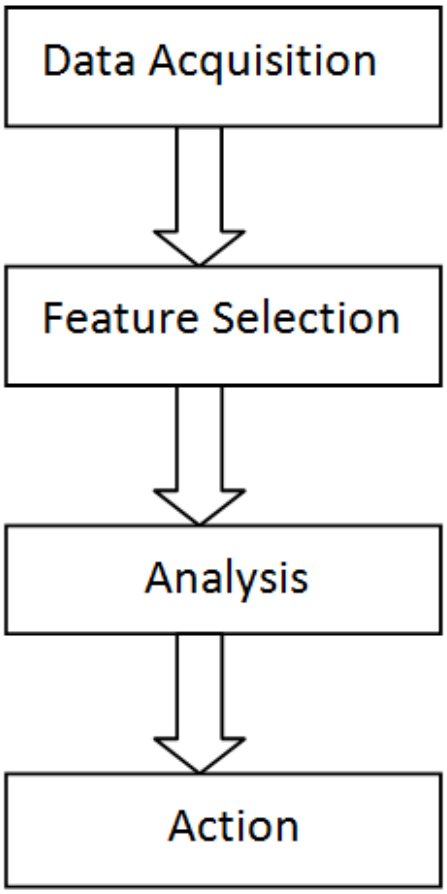

Fig. 1. Working of Intrusion Detection System

C. Performance Measurement of IDS [17] [19]

There are some primary factors which are used during performance measurement of Intrusion detection system.
True positive (TP): The total number of normal data which are detected as a normal data during intrusion detection process.

True negative (TN): In Intrusion detection, number of detected abnormal data which are actually abnormal data in dataset.

False positive (FP): Or false alarm, total number of detected normal data but they are actual attack.

False negative (FN): Number of detected abnormal instances but in real they are normal data.

Performance of IDS is measured in terms of detection rate, accuracy and false alarm rate.

Detection Rate (DR) = (TP/TP+FN) x 100% *False Alarm Rate (FAR)* =FP/Number of Attacks *Accuracy* = (TP+TN/TP + TN + FP + FN)x 100%

IV. DATA MINING BASED INTRUSION DETECTION SYSTEM

Data mining is the activity of extracting relevant information from a large amount of data.[20]

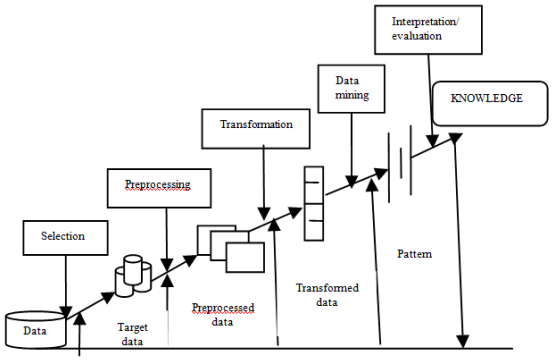

Fig. 2. Data Mining

Network traffic is massive and information comes from different sources, so the dataset for IDS becomes large. Hence the analysis of data is very shard in case of large dataset. Data mining techniques are applied on IDS because it can extract the hidden information and deals with large dataset. Presently Data mining techniques plays a vital role in IDS. By using Data mining techniques, IDS helps to detect abnormal and normal patterns.

This section describes different Data mining techniques such as clustering and classification, which are used in IDS to obtain information about vulnerability by monitoring network data.[1]

A. Classification [1]

Classification is the task of taking each and every instances of dataset under consideration and assigning it to a particular class normal and abnormal means known structure is used for new instances. It can be effective for both misuse detection and anomaly detection, but more frequently used for misuse detection. Classification categorized the datasets into predetermined sets. It is less efficient in intrusion detection as compared to clustering. Different classification techniques such as decision tree, naive bayes classifier, K-nearest neighbour classifier, Support vector machine etc are used in IDS.

1) Decision Tree [21]

Decision tree is a recursive and tree like structure for expressing classification rules. It uses divide and conquer method for splitting according to attribute values. Classification of the data proceeds from root node to leaf node, where each node represents the attribute and its value & each leaf node represent class label of data. Tree based classifier have highest performance in case of large dataset. Different decision tree algorithms are described below[6]

ID3 algorithm

It is famous decision tree algorithm developed by Quinlan. ID3 algorithm basically attribute based algorithm that constructs decision tree according to training dataset. The attribute which has highest information gain is used as a root of the tree.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

J48 algorithm
It is based on ID3 algorithm and developed by Ross Quinlan. In WEKA, C4.5 decision tree algorithm is known as J48 algorithm. It construct decision tree using information gain, attribute which have highest information gain is selected to make decision. The main disadvantage of this algorithm is that it takes more CPU time and memory in execution. Another different tree based classifier [15]:

AD Tree
Alternating decision tree is used for classification. AD Tree have prediction node as both leaf node and root node.

NB Tree
NB Tree algorithm uses both decision tree and naive bayes classifier. Root node uses decision tree classifier and leaf nodes uses naive bayes classifier.

Random Forest [22]
Random Forest is first introduced by Lepetit et.al. and it is ensemble classification technique which consists of two or more decision trees. In Random Forest, every tree is prepared by randomly select the data from dataset. By using Random Forest improve the accuracy and prediction power because it is less sensitive to outlier data. It can easily deals with high dimensional data.

2) K-Nearest Neighbor [21]
It is one of the simplest classification technique. It calculates the distance between different data points on the input vectors and assigns the unlabeled data point to its nearest neighbor class. K is an important parameter. If k=1, then the object is assigned to the class of its nearest neighbor. When value of K is large, then it takes large time for prediction and influence the accuracy by reduces the effect of noise.

3) Naive Bayes classifier [13]
Naive Bayes classifier is probabilistic classifier. It predicts the class according to membership probability. To derive conditional probability, it analyzes the relation between independent and dependent variable.
Bayes Theorem:

$$P(H/X)=P(X/H).P(H)/P(X)$$

Where, X is the data record and H is hypothesis which represents data X and belongs to class C. P(H) is the prior probability, P(H/X) is the posterior probability of H conditioned on X and P(X/H) is the posterior probability of X conditioned on H.

Construction of Naive Bayes is easy without any complicated iterative parameter. It may be applied to large number of data points but time complexity increases.

4) Support Vector Machine [8]
Support Vector Machine is supervised learning method used for prediction and classification. It separate data points into two classes +1 and -1 using hyperplane because it is binary classification classifier. +1 represents normal data and -1 for suspicious data.
Hyperplane can be expressed as: W. X +b=0

Where $W=\{w_1,w_2,.....,w_n\}$ are weight vector for n attributes $A=\{A_1,A_2,.....,A_n\}$, $X=\{x_1,x_2,.....,x_n\}$ are attribute values and b is a scalar. The main goal of SVM is to find a linear optimal hyper plane so that the margin of separation between the two classes is maximized. The SVM uses a portion of the data to train the system.

B. Clustering [1]
Since the network data is too huge, labelling of each and every instances or data points in classification is expensive and time consuming. Clustering is the technique of labelling data and assign into groups of similar objects without using known structure of data points. Members of same cluster are similar and instances of different clusters are different from each other. Clustering technique can be classified into four groups: Hierarchical algorithm, Partitioning algorithm, Grid based algorithm and Density based algorithm. Some clustering algorithms are explained here.

1) K-Means Clustering algorithm [23][13]
K-Means clustering algorithm is simplest and widely used clustering technique proposed by James Macqueen. In this algorithm, number of clusters K is specified by user means classifies instances into predefined number of cluster. The first step of K-Means clustering is to choose k instances as a center of clusters. Next assign each instances of dataset to nearest cluster. For instance assignment, measure the distance between centroid and each instances using Euclidean distance and according to minimum distance assign each and every data points into cluster. K –Means algorithm takes less execution time, when it applied on small dataset. When the data point increases to maximum then it takes maximum execution time. It is fast iterative algorithm but it is sensitive to outlier and noise.

2) K-Medoids clustering algorithm [13]
K-Medoids is clustering by partitioning algorithm as like as K-means algorithm. The most centrally situated instance in a cluster is considered as centroid in place of taking mean value of the objects in K-Means clustering. This centrally located object is called reference point and medoid. It minimizes the distance between centroid and data spoints means minimize the squared error. K-Medoids algorithm performs better than K-Means algorithm when the number of data points increases to maximum. It is robust in presence of noise and outlier because medoid is less influenced by outliers, but processing is more expensive.

V. CONCLUSION

On the basis of detection rate, accuracy, execution time and false alarm rate, the paper has analyzed different classification and clustering data mining techniques for intrusion detection. According to given necessary parameter, execution time of Support vector machine is less and produces high accuracy with smaller dataset, while construction of Naive Bayes classifier is easy. Also decision tree has high detection rate in case of large dataset. In clustering techniques, execution time of K-Means clustering algorithm is less in case of small dataset, but when number of data point increases, K-Medoids performs better.

REFERENCES

[1] Deepthy K Denatious & Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection", International Conference on Computer Communication and Informatics (ICCCI - 2012), Jan. 10 – 12, 2012, Coimbatore, INDIA

[2] Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, "Using Rough Set And Support Vector Machine For Network Intrusion Detection", International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, April 2009

[3] Deepak Upadhyaya and Shubha Jain, "Hybrid Approach for Network Intrusion Detection System Using K-Medoid Clustering and Naïve Bayes Classification", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 1, pp 231-236, May 2013

[4] Xiang, M.Y. Chong and H. L. Zhu, "Design of Multiple-level Tree classifiers for intrusion detection system", IEEE conference on Cybernetics and Intelligent system, 2004

[5] Peddabachigiri S., A. Abraham., C. Grosan and J. Thomas, "Modeling of Intrusion Detection System Using Hybrid intelligent systems" , Journals of network computer application, 2007

[6] Mrutyunjaya Panda and Manas Ranjan Patra, "A Comparative Study Of Data Mining Algorithms For Network Intrusion Detection", First International Conference on Emerging Trends in Engineering and Technology, pp 504-507, IEEE, 2008

[7] M.Govindarajan and Rlvl.Chandrasekaran, "Intrusion Detection Using k-Nearest Neighbor" pp 13-20, ICAC, IEEE, 2009

[8] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi and Lilly Suriani Affendey, "Intrusion Detection Using Data Mining Techniques", pp 200-203, IEEE, 2010

[9] Song Naiping and Zhou Genyuan, "A study on Intrusion Detection Based on Data Mining", International Conference of Information Science and Management Engineering , Pp 135- 138, IEEE,2010

[10] T. Velmurugan and T. Santhanam, "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points", Journal of Computer Science 6 (3): 363-368, 2010

[11] P Amudha and H Abdul Rauf, "Performance Analysis of Data Mining Approaches in Intrusion Detection", IEEE, 2011

[12] R.China Appala Naidu and P.S.Avadhani, "A Comparison of Data Mining Techniques for Intrusion Detection", International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp-41-44, IEEE, 2012

[13] Roshan Chitrakar and Huang Chuanhe, "Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes Classification", IEEE,2012

[14] Roshan Chitrakar and Huang Chuanhe, "Anomaly Detection using Support Vector Machine Classification with k-Medoids Clustering", IEEE, 2012

[15] Sumaiya Thaseen and Ch. Aswani Kumar, "An Analysis of Supervised Tree Based Classifiers for Intrusion Detection System", International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), IEEE, February 21-22 2013

[16] David Ndumiyana, Richard Gotora and Hilton Chikwiriro, "Data Mining Techniques in Intrusion Detection: Tightening Network Security", International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 5, May – 2013

[17] Muhammad K. Asif, Talha A. Khan,Talha A. Taj, Umar Naeem and Sufyan Yakoob, " Network Intrusion Detection and its Strategic Importance", Business Engineering and Industrial Applications Colloquium(BEIAC), IEEE, 2013

[18] Kapil Wankhade, Sadia Patka and Ravindra Thools, "An Efficient Approach for Intrusion Detection Using Data Mining Methods", IEEE 2013

[19] Fatin Norsyafawati Mohd Sabri, Norita Md Norwawi and Kamaruzzaman Seman, "Hybrid of Rough Set Theory and Artificial Immune Recognition System as a Solution to Decrease False Alarm Rate in Intrusion Detection System", IEEE 2011

[20] Vaishali B Kosamkar and Sangita S Chaudhari, "Data Mining Algorithms for Intrusion Detection System: An Overview", International Conference in Recent Trends in Information Technology and Computer Science (ICRTITCS), 2012

[21] Hind Tribak , Blanca L. Delgado-Marquez, P.Rojas, O.Valenzuela, H. Pomares and I. Rojas, " Statistical Analysis of Different Artificial Intelligent Techniques applied to Intrusion Detection System", IEEE, 2012

[22] S. Revathi and A. Malathi, "Data Preprocessing for Intrusion Detection System using Swarm Intelligence Techniques", International Journal of Computer Applications , Volume 75– No.6, August 2013

[23] Iwan Syarif, Adam Pruge Bennett and Gary Wills, "Unsupervised clustering approach for network anomaly detection", IEEE.