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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

An Overview of Data Mining and Anomaly Intrusion Detection System using K-Means

S.Sujatha[1] , P.Hemalatha[2] ,S.Devipriya[3]

Assistant Professor, Department of Computer Science, Sri Akilandeswari Women's College, Vandavasi, India

P.G. Student, Department of Computer Science, Sri Akilandeswari Women's College, Vandavasi, India

P.G. Student, Department of Computer Science, Sri Akilandeswari Women's College, Vandavasi, India

ABSTRACT: In Information Security, intrusion detection is the act of detecting actions that attempt to compromise the security goal. One of the primary challenge to intrusion detection is the problem of misjudgement , misdetection and lack of real time response to the attack. Although various techniques or applications are available to protect data, loopholes exist. Thus to analyze data and to determine various kind of attack data mining techniques have emerged to make it less vulnerable. Anomaly detection uses these datamining techniques to detect the surprising behaviour hidden within data increasing the chances of being intruded or attacked. Various data mining techniques as clustering, classification and association rule discovery are being used for intrusion detection. The proposed technique combines data mining approaches like K Means clustering algorithm and RBF kernel function of Support Vector Machine as a classification modules. The main intention of proposed technique is to decrease the number of attributes associated with each data point. So, the future technique can perform better in terms of Detection Rate and Accuracy when applied to KDDCUP'99 Data Set. This paper reviews various data mining techniques for anomaly detection to provide better understanding among the existing techniques that may help interested researchers to work future in this direction.

KEYWORDS: Data Mining; Anomaly Detection, K- means Clustering, Classification, Intrusion Detection System, KDD data set.

I.INTRODUCTION

Intrusion Detection Systems (IDS) are security tools that provided to strengthen the security of communication and information systems. This move towards is similar to other measures such as antivirus software, firewalls and access control schemes. Conventionally, these systems have been classified as a signature detection system, an anomaly detection system. In signature based detection, the system identifies patterns of traffic or application data is presumed to be malicious while anomaly detection systems compare activities against a normal defined behaviour. An intrusion detection system is a type of security management for computers and networks. Some preventive systems are able to detect attacks in real-time and can stop an attack. Other systems are designed to audit some fruitful information about attacks. Such systems can help to detect such attack and reduce the possibility of such attack in future. An IDS continuously monitors the network or a computer system for any suspicious or malicious activity and as soon as it detects any such kind of activity, analyses it and alarms the system or network administrator. It becomes very difficult for an attacker to carry out any activity without setting bad an alarm. Finally, it can detect the attacks that are previously not known. Anomaly detection systems look for anomalous events rather than the attacks.

A. Anomaly Detection

Anomaly detection is the process of finding the patterns in a dataset whose behavior is not normal on expected. These unexpected behaviors are also termed as anomalies or outliers. The anomalies cannot always be categorized as an attack but it can be a surprising behavior which is earlier not known. It may or may not be harmful. The anomaly detection provides very significant and critical information in various applications. When data has to be analyzed in order to find relationship or to predict known or unknown data mining technique are used. These include clustering,

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

classification and machine based learning techniques. Hybrid approaches are also being created in order to attain higher level of accuracy on detecting anomalies. In this advance the authors try to combine existing data mining algorithms to derive recovered results. Thus detecting the unusual or unexpected behavior or anomalies will yield to study and categorize it into new type of attacks or any particular type of intrusions. This study attempts to provide a better understanding among the various types of data mining approaches towards anomaly detection that has been made until now.

B. Basic Methodology of anomaly detection technique

Even though different anomaly approaches exists, as shown in figure 1 parameter wise train a model prior to detection.
Parameterization: Pre processing data into a pre-recognized format such that it is suitable or in accordance with the targeted systems behavior.

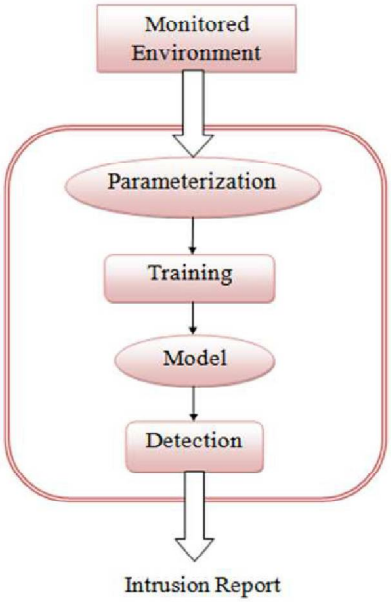

Fig1. Methodology of Anomaly Detection

Training stage: A model is built on the basis of normal (or abnormal) behavior of the system. There are different ways that can be opted depending on the type of anomaly detection measured. It can be both manual and automatic.

Detection stage: When the model for the system is available, it is compared with the (parameterized or the pre defined) observed traffic. If the deviation found exceeds (or is less than when in the case of abnormality models) from a pre defined threshold then an alarm will be triggered.

II. ANOMALY DETECTION USING DATA MINING TECHNIQUES

Anomalies are pattern in the data that do not conform to a well defined normal behavior. The cause of anomaly may be a malicious activity or some kind of intrusion. This unusual behavior found in the dataset is interesting to the analyst and this is the most important feature for anomaly detection [14]. Anomaly detection is a topic that had been covered under various survey, review articles and books [4, 5]. Phua et al (2010) have done a detailed survey on various fraud detection techniques that has been carried out in the past few years. They have defined the professional fraudster, the

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

main types and subtypes of known fraud, and also presented the nature of data evidence collected within affected industries [6]. Padhy et al (2012) provide a detailed survey of data mining applications and its feature scope.
They fixed that anomaly detection is an application of data mining where various data mining techniques can be applied [3]

In this paper review of different approaches of anomaly detection focuses on the broad classification of existing data mining techniques. Data mining consists of four classes of task; they are association rule learning, clustering, classification and regression. Next subsection presents anomaly detection techniques under these four classes of task:

A. Clustering based Anomaly Detection techniques

Clustering can be defined as a division of data into group of similar objects. Each group, or bunch, consists of objects that are similar to one another and dissimilar to objects in other group [13]. Clustering algorithms are able to detect intrusions without previous knowledge. There are various methods to perform clustering that can be applied for the anomaly discovery. Following is the description of some of the proposed approaches

k-Means: k-Means clustering is a cluster analysis method where we define k disjoint clusters on the basis of the feature value of the objects to be group. Here, k is the user defined parameter [9]. There has been a Network Data Mining (NDM) approach which deploys the K-mean clustering algorithm in order to separate time intervals with normal and anomalous traffic in the guidance dataset. The resulting cluster centroids are then used for fast anomaly detection in monitoring of new data [10].

k-Medoids: This algorithm is very similar to the k-Means algorithm. It differs mainly in its representation of the different cluster. Here each cluster is represented by the most centric object in the cluster, rather than by the implicit mean that may not belong to the bunch. The k-medoids method is more robust than the k-means algorithm in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean.This method detects network anomalies which contains unknown intrusion. It has been compared with various other clustering algorithms and have been find out that when it comes to accuracy, it produces much better results than k-Means [11].

B. Classification based anomaly detection

Classification can be defined as a problem of identifying the category of new instances on the basis of a training set of data containing observations (or instances or tuples) whose category partisanship is known. The category can be termed as class label. Various instances can belong to one or many of the class label. In machine learning, classification is considered as an instance of supervised learning for example learning where a training set of correctly-identified observations is available. An algorithm that implements classification is known as a classifier. It is constructed to predict categorical labels or class label quality. In case of
anomaly detection it will classify the data generally into two categories namely normal or abnormal. Following are common machine learning technologies in anomaly detection

Classification Tree: In machine learning classification tree is also called as a prediction model or decision tree. It is a tree pattern graph which is similar to flow chart structure; the internal nodes are a test property, each branch represents test result, and final nodes or leaves represent the class to which any object belongs. The most fundamental and common algorithm used for classification tree is ID3 and C4.5 There are two methods for tree construction, topdown tree construction and bottom-up pruning. ID3 and C4.5 belong to top-down tree construction [16]. Further classification tree approaches when compared to naïve bayes classification, the result obtained from decision trees was found to be more accurate [19]
Support Vector Machine: These are a set of related supervised learning methods used for classification and regression. Support Vector Machine (SVM) is widely applied to the field of pattern recognition. It is also used for an intrusion detection structure. The one class SVM is base on one set of examples belonging to a particular class and no negative examples rather than positive examples . When compared to neural network in KDD cup data set, it was found out that SVM out performed NN in terms of false alarm rate and accuracy in most kind of attacks [18].

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

III. RELATED WORK

Data mining technology to Intrusion Detection Systems can mine the features of new and unknown attacks well, which is a maximal help to the Intrusion Detection System. This work is performed on KDD Cup 99 data set to analyze the effectiveness between our planned method and the conventional algorithms. The presentation of the various algorithm measured in terms of accuracy, detection rate and false alarm rate. The best possible accuracy and detection rate can be achieved by using our proposed hybrid learning approach. Different classifiers can be use to formed a hybrid learning approaches such as combination of clustering and classification technique. In 2009, Meng Jianliang, Shang Haikun, Bian Ling had presented the K-means algorithm [3] for intrusion detection. Experimental results on a subset of KDD dataset showed that the detection rate stayed always above 96% while the false alarm rate was below 2%.The time complexity is low

IV. PROPOSED WORK

The literature survey represents various hybrid techniques for intrusion detection. Each technique has its own benefits and their own shortcomings. As well as performance of each technique is varies in terms of Accuracy, Detection rate & False Positive Rate. The proposed technique combines unsupervised learning with supervised learning [12]. For the first stage in proposed hybrid learning approach, we group similar data instances based on their behaviours by utilizing K Means Clustering as a pre classification component. Next using RBF kernel function of SVM classifier we classify the clusters into attack classes as a final task. We found that the data during misclassified during earlier stage may be correctly classified in the subsequent classification stage.

A. *K Means Algorithm for Intrusion Detection:*

1. Choose randomly five data records as initial Clusters Mean (cluster centre).
2. Calculate the new centriod for the dataset, for each data record x from D,
3. Calculate the Euclidean distance between data record x and each cluster mean.
4. Assign data record x to the closest cluster.
5. Re-calculate the mean for current cluster collections.
6. Repeat the procedure until we get stable clusters.
7. Use these centriods classification of anomaly and normal traffic.

The objective function is [7]:

$J = \sum_{i=1}^{k} \sum_{j=1}^{n} dij(xj,ci)$

Where dij(Xj,Ci) is a chosen distance measure (Euclidean distance) between a data point xj and the cluster center ci, is an indicator of the distance of the data points from their respective cluster centers.

V. EXPERIMENTS & RESULTS

To evaluate the effectiveness of proposed approach we had used KDDCUP 99 Dataset. From these dataset we had created small training dataset as well as testing dataset for each attack class. These dataset we had used for experimental purpose. The results we had calculated in terms of Detection Rate (DR) and Accuracy (ACC) [11].

Each metric is defined below:

• Detection Rate (DR): Detection rate is the rate of correctly classified intrusive examples to the total number of intrusive examples.

Detection Rate (DR) = (TP) / (TP+FP)

• Accuracy (ACC): Accuracy [2] is the ratio of correctly classified to the total classified examples.

Accuracy (ACC) = (TP+TN) / (TP+TN+FP+FN)

Where

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

FN is False Negative,
TN is True Negative,
TP is True Positive and
FP is False Positive.

Following table II shows the accuracy results for all attack category classes obtained from K- Means (KM), SVM classifier & proposed system K-Means with RBF Kernel function. This proposed system we had renamed as KMSVM. The dataset used contains all 41 attributes. In above table we can see accuracy of our proposed KMSVM is 92.86% whereas KM has accuracy 86.67% and accuracy of SVM is 40% for DOS attack. Thus, we can observe
that KMSVM performs better than K-Means & SVM classifier. In the same way we can observe that for all other attacks like for PROBE, U2R and R2L proposed KMSVM performs better compared to others.

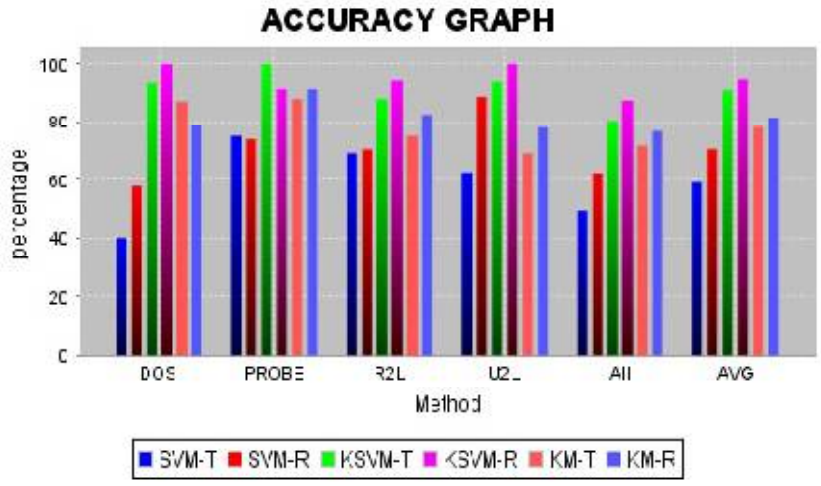

Fig2 .Accuracy Graph

VI. CONCLUSION

In this paper various data mining techniques are described for the anomaly detection that had been proposed in the past few years. This review will be helpful to researchers for gaining a basic insight of various approaches for the anomaly detection. Although much work had been done using independent algorithms, hybrid approaches are being vastly used as they provide better results and overcome the drawback of one approach over the previous. Every day new unknown attacks are witnessed and thus there is a need of those approaches that can detect the unknown behaviour in the data set store, transferred or modified. In this research work fusion or combination of already existing algorithms are mentioned that have been proposed. Intrusion detection systems play an important role in network security. Feature selection is the major challenging issue in IDS in order to reduce the useless and redundant features among the attributes. In this report, an hybrid learning approach through combination of K - Means clustering and SVM classifier are proposed. In hybrid IDS we have used RBF kernel function of SVM for classification purpose. We took the help of K- Means clustering technique to reduce large heterogeneous dataset to a number of small homogeneous subsets. The proposed approach is compared and evaluated using KDD CUP 99 dataset. As well as the false alarm rate also decreases of proposed technique.

REFERENCES

[1] Meng Jianliang, Shang Haikun Bian Ling. The application on intrusion detection based on k-means cluster algorithm. In International Forum on Information Technology and Application, pages 150–152. IEEE,2009.
[2] Meijuan Gao, Jingwen Tian, Fan Zhang. Network intrusion detection method based on radial basic function neural network. In Computer Engineering and Design, vol. 29,no. 12, pages 42–46. IEEE, 2009.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

[3] Preecha Somwang, Woraphon Lilakiatsakun. Computer network security based on support vector machine approach. In 11th International Conference on Control, Automation and Systems.
[4] XIE Yang, ZHANG Yilai. An intelligent anomaly analysis for intrusion detection based on svm. In International Conference on Computer Science and Information Processing (CSIP), pages 739–742. IEEE, 2012.
[5] Susheel Kumar Tiwari, Sanjay Kumar Sharma, Pankaj Pande, Mahendra Singh Sisodia. An improved network intrusion detection technique based on k-means clustering via naïve bayes classification. In International Conference On Advances In Engineering, Science And Management (ICAESM -2012).
[6] Roshan Chitrakar, Huang Chuanhe. Anomaly detection using support vector machine classification with k-medoids clustering. In computer society symposium on research in security and privacy, pages 46–50. IEEE, 2012.
[7] Meng Jianliang, Shang Haikun Bian Ling,. The application on intrusion detection based on k-means cluster algorithm. In International Forum on Information Technology and Application pages 150–152. IEEE, 2009.
[8] Mr. V. K. Pachghare, Parag Kulkarni. Pattern based network security using decision trees and support vector machine. pages 254–257. IEEE,2011.
[9] Syarif I., Prugel-Bennett A., Wills G., Data mining approaches for network intrusion detection from dimensionality reduction to misuse and anomaly detection; Journal of Information Technology Review ; 3(2); 2012; p. 70-83.
[10] Han J., Kamber M., Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.
[11] Berkhin P., A survey of clustering data mining techniques;Grouping multidimensional data; Springer Berlin Heidelberg; 2006; p. 25-71.
[12] Dokas P., . Ertoz L., Kumar V., Lazarevic A., Srivastava J., Tan P. N., Data mining for network intrusion detection, In Proceedings of NSF Workshop on Next Generation Data Mining; 2002; p. 21-30
[13] Garcia-Teodoro P., Diaz-Verdejo J., Maciá-Fernández G. , Vázquez E., Anomaly-based network intrusion detection: Techniques, systems and challenges; Computers and security; 28(1); 2009; p. 18-28.
[14] Wu S. Y., Yen E., Data mining-based intrusion detectors; Expert Systems with Applications; 36(3); 2009 p. 5605-5612.
[15] Kaur N., Survey paper on Data Mining techniques of Intrusion Detection;International Journal of Science, Engineering and Technology Research; 2(4); 2013; p. 799-804.
[16] Tang D. H., Cao Z.,Machine Learning-based Intrusion Detection Algorithm; Journal of Computational Information Systems;5(6); 2009; p. 1825-1831.
[17] Amor N. B., Benferhat S., Elouedi Z., Naive Bayes vs decision trees in intrusion detection systems, In Proceedings of the ACM symposium on Applied computing; 2004; p. 420-424
[18] Kou Y., Lu C. T., Sirwongwattana S., Huang Y. P., Survey of fraud detection techniques; In Proceedings of the IEEE International conference Networking, sensing and control; 2; 2004; p. 749-754.
[19] TsaiC. F., Hsu Y. F., Lin C. Y., Lin W. Y. , Intrusion detection by machine learning: A review; Expert Systems with Applications; 36(10); 2009; p. 11994- 12000.
[20] Farid D. M., Harbi N., Rahman M. Z., Combining naive bayes and decision tree for adaptive intrusion detection; International Journal of Network Security & Its Applications (IJNSA);2(2);2010;p. 12-25.