COMPARISON OF DATA MINING TECHNIQUES USED IN ANOMALY BASED IDS

A.M.CHANDRASHEKHAR[1], CHANDANA P[2]

[1] Assistant Professor, Department of Computer Science & Engineering, Sri Jayachamarajendra College of Engineering(SJCE), JSS S&T University Campus, Mysore, Karnataka, India
[2] M.Tech 2nd Semester, Department of Computer Science & Engineering, Sri Jayachamarajendra College of Engineering(SJCE), JSS S&T University Campus, Mysore, Karnataka, India

---***---

Abstract - *The fundamental idea of intrusion detection system is to identify the attacks against information present in the system. Misuse based and anomaly based detection system are the categories of intrusion detection system. This paper introduces the intrusion detection system and its types. It analysis about the anomaly based system and its applications in various fields. This review contribute a better perception about the data mining techniques in anomaly detection system.*

Key Words: **Intrusion detection, anomaly, network, attacks and classification tree.**

1. INTRODUCTION

Intrusion detection system (IDS) is a system which determines attacks in the system. It is a type of security system for computer and networks. It identifies possible rifts in networks. The activities carried out by intrusion detection system are supervising and evaluating the user and system actions, capability to discover the patterns of attacks, analysis of abnormal activity patterns and tracking user policy violations. The individual packets flowing through the network are analyzed using network based systems. Each individual computer or host is analyzed using host based system. Network intrusion system consists of two types like misuse based and anomaly based system. It retains database of previous attacks and compare when found any attack in a system. It shows various data mining techniques in anomaly based intrusion detection system.

2. CATEGORIES OF IDS

IDS can be classified in two broad categories: Misuse detection and Anomaly detection.

Misuse Detection:

The system learns patterns from already known attacks. These accomplished patterns are examined through the incoming data to find intrusions of the already known types. This way is not capable in detecting new attacks that do not replace pre-defined patterns. Notice a security guard present at an entrance who is responsible for allowing only valid persons to pass through the gate. One way that the guard may maintain a database of photographs of well-known fugitive who should not be allowed entry. The guard can then analyze each incoming person with the database and find out if the person is one of those fugitive. If so, the guard prevents the fugitive from passing through the entrance. The complexity here is that a fugitive whose photograph is not in the database entrance. The difficulty here is that a fugitive whose photograph is not in the database will be allowed entry. This approach corresponds to the Misuse Detection technique.

Anomaly Detection:

In this type patterns are accomplished from normal data. The invisible data is evaluated and examined to find deviations from these learned patterns. [1] These breaches are 'anomalies' or possible intrusions. This type is not efficient for recognizing the form of attack. In this way the escort may follow is to maintain a database of photographs of all the valid persons to be allowed entry. The escort allows entry to the incoming person, only if his photograph is erect in the database. This way, all persons whose photographs are not erect in the database are identified as culprits and not allowed entry. This idea corresponds to the Anomaly Detection technique. To grab the profits of both the approaches, an IDS system should associate anomaly detection and misuse detection techniques. [7] Data mining from intrusion detection point of view is the finding of malicious activity patterns or normal activity patterns from the huge amount of data conveying through the network or stored in system logs. The data on which data mining is skilled should contain all possible forms of normal data. The data should be wealthy enough so that no normal data is miss-interpreted as an anomaly.

3. METHODOLOGY

Basic Methodology of anomaly detection technique is explained below. Although different anomaly approaches exists, the simple steps used are as shown in figure 1 which is parameter wise training a model prior to detection.

The main 3 stages are: Parameterization, training stage and detection stage.

Parameterization: Pre-processing data into a pre-established formats such that it is acceptable or in consonance with the targeted systems behavior.

Training stage: A model is built on the basis of normal or abnormal behavior of the system. [8] There are different ways that can be preferred depending on the type of anomaly detection considered. It can be both manual and automatic.

Detection stage: When the model for the system is available, it is compared with the recognized 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.

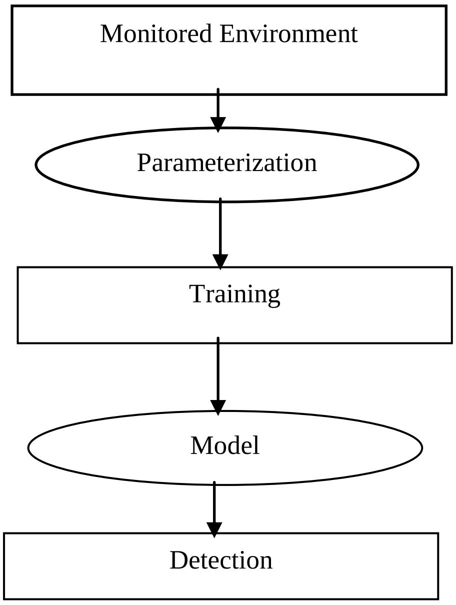

Fig 1: Methodology of anomaly detection

Using the parameters and environment, training is given, modeled and then detected with recognized traffic. It is checked for any deviations in traffic of network.

4. ANOMALY DETECTION USING DATA MINING TECHNIQUES

Anomalies are pattern in the data that do not accommodate to a well describe normal behavior. [9] The element of anomaly may be a malicious activity or some kind of intrusion. This abnormal behavior found in the dataset is interesting to the analyst and this is the most important component for anomaly detection. Some of the scholars stated that anomaly detection is an application of data mining where various data mining techniques can be applied. They described readymade data mining techniques that can be applied directly to detect the intrusion. They contribute a wide prospective to the techniques that they can be practically deployed by viewing the possible purposes for the lack of acceptance to the proposed novel approaches.

The statistical based anomaly detection technique:

The statistical based anomaly detection techniques clarify the issue with string, logic and rule based anomaly detection. [2] Withal, the current statistic in HIDES/NIDES are univariate technique, aids that it is applied to only one behavior measure, while countless intrusions contains multiple subject and plenty events having effect on multiple behavior measures. Thus a multivariate anomaly detection technique is required for intrusion detection. So for this reason, there are plenty multivariate techniques are used to examine and investigate anomaly in manufacturing systems. Some of them are Hoteling, multivariate cumulative sum, and multivariate exponentially weighted moving average. According to theoretical details these multivariate statistical forms can be imposed to intrusion detection for examining and detecting anomaly of a subject in the field of information science. According to practical values it is not feasible because of the computationally comprehensive forms of these statistical techniques and they cannot accommodate the requirements of intrusion detection systems for huge reasons. First, intrusion detection systems accord with enormous amount of high-dimensional process data because of huge number of behaviors and a high frequency of subject's events occurrence. Second, intrusion detection systems claimed a minimum delay of processing of each fact in computer systems to make sure an early detection and signals of intrusions. Therefore, a multivariate anomaly detection type with low computation cost is Chi-Square statistic, which is good candidate for intrusion detection. Chi square perform as multivariate but it has characteristic of robustness, so it affected above problems.

Clustering based Anomaly Detection techniques are: k-Means, k-Medoids, EM clustering, Outlier detection algorithms.

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 grouped. [3] Here, k is the user defined parameter. There has been a Network Data Mining (NDM) way which uses the K-mean clustering algorithm in order to separate time intervals with normal and anomalous traffic in the training dataset. The produced cluster centroids are then used for fast anomaly detection in monitoring of new data.

k-Medoids: This algorithm is very similar to the k-Means algorithm. It differs mainly in its representation of the different clusters. 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 cluster. 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.

EM Clustering: This algorithm can be explored as an extension of k means which assigns an object to the cluster to which it is similar, based on the mean of cluster. [10] In this way instead of assigning object in the dedicated cluster, assign the object to a cluster according to a weight representing the probability of membership. In other words there are no strict lines in between the clusters. Here new mean is computed on the support of weight measures. When contrasted to k means and k medoids, EM outperformed them and resulted in higher accuracy.

Outlier Detection Algorithms: Outlier detection is a technique to find patterns in data that do not conform to expected behavior. Since an outlier can be defined as a data point which is very different from the rest of the data, based on certain measures. There are several outlier detection schemes. User can select any one of them on the basis of its efficiency and how he can deal the problem of anomaly detection. One of the approach is Distance based Approach. It is based on the Nearest Neighbor algorithm and implements a well-defined distance metric to detect outliers. Greater the distance of the object to its neighbor, the more likely it is to be an outlier. It is an efficient approach in detecting probing attacks like Denial of Service (DoS) attacks. Other one is Density based local outlier approach. Distance based outlier detection confide on the overall or global distribution of the given set of data points. [11] The data is not uniformly distributed thus the distance based way encounter various difficulties during analysis of data. The main belief of this density based type is to assign to each data example a degree of being outlier, which is called the Local Outlier Factor (LOF). The outlier factor is regional in the sense that only a restricted neighborhood of each object is considered. Various other algorithms are proposed for anomaly detection in the Wireless Sensor Networks (WSN). A hierarchical framework have been proposed to overcome challenges in WSN's where an accurate model and the approximated model is made learned at the remote server and sink nodes. A roughed local outlier factor algorithm is also proposed which can be learned at the sink nodes for the detection model in WSN. These afford more efficient and accurate results.

Classification based anomaly detection techniques are: Classification tree, Fuzzy logic, Naïve Bayes network, Genetic algorithm, Neural networks.

Classification Tree: In machine learning classification tree is also called as a prediction model or decision tree. [4] 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. There are two methods for tree construction, top down tree construction and bottom-up pruning. ID3 and C4 belong to top-down tree construction. Further classification tree approaches when compared to naïve Bayes classification, the result obtained from decision trees was found to be more accurate.

Fuzzy Logic: It is borrowed from fuzzy set theory which accord with reasoning that is rough value rather than precisely deduced from classical predicate logic. [12] The application side of fuzzy set theory accord with well thought out real world expert values for a complex problem. In this approach the data is classified on the basis of various statistical metrics. These portions of data are applied with fuzzy logic rules to classify them as normal or malicious. There are many other fuzzy data mining techniques to excerpt patterns that represent normal behavior for intrusion detection that describe a diversity of modifications in the existing data mining algorithms in order to increase the efficiency and accuracy .

Naïve Bayes network: There are many cases where the statistical dependencies or the causal relationships between system variables exist. [5] It can be difficult to precisely express the probabilistic relationships among these variables. In other words, the former knowledge about the system is simply that some variable might be influenced by others. To take advantage of this structural relationship between the random variables of a problem, a probabilistic graph model called Naïve Bayesian Networks (NB) can be used. This model provides answer to the questions like if few observed events are given then what is the probability of a particular kind of attack. It can be done by using formula for conditional probability. [13] The structure of a NB is typically represented by a Directed Acyclic Graph (DAG) where each node represents one of system variables and each link encodes the influence of one node upon another. When decision tree and Bayesian techniques are compared, though the accuracy of decision tree is far better but computational time of Bayesian network is low. Hence, when the data set is very large it will be efficient to use NB models.

Genetic Algorithm: It was popularized in the field of computational biology. These algorithms reside in the larger class of Evolutionary Algorithms (EA). [14] They evolve solutions to optimization problems using routines influenced by natural evolution, such as inheritance, selection, mutation and crossover. Since then, they have been enforced in various fields with very promising results. In intrusion detection, the Genetic Algorithm (GA) is applied to derive a set of classification rules from the network audit data. The support-confidence framework is utilized as a fitness function to judge the quality of each rule. Significant properties of GA are its robustness against noise and self-learning capabilities. The uses of GA techniques reported in case of anomaly detection are high attack detection rate and lower false-positive rate.

Neural Networks: It is a set of interconnected nodes designed to imitate the functioning of the human brain. [15] Each node has a weighted connection to several other nodes in neighboring layers. Particular nodes take the input received from connected nodes and value the weights together with a simple function to compute output values. Neural networks can be assured for supervised or unsupervised learning. The user points out the number of hidden layers as well as the number of nodes within a specific hidden layer. Depending on the application, the output layer of the neural network may contain one or several nodes. The Multilayer Perceptions (MLP) neural networks have been very successful in a range of applications and producing more accurate results than other existing computational learning models. [16] They are capable of approximating to random accuracy, any continuous function as long as they contain enough hidden units. This means that such models can form any classification decision boundary in feature space and thus act as non-linear discriminate function. 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 system. The one class SVM is placed on one set of examples belonging to a particular class and no negative illustrations rather than using positive and negative example. When correlated to neural networks in KDD cup data set, it was erect out that SVM out performed NN in terms of false alarm rate and accuracy in most kind of attacks.

Hybrid approaches:

Cascading supervised techniques: Here various classification algorithms are merged together in order to obtain higher accuracy. [6] A combination of naïve Bayes and decision tree algorithm was proposed. This hybrid algorithm was evaluated in Knowledge Data Discovery (KDD) cup dataset and the accuracy achieved was 99 percent. It focuses on the development of the performance of Naïve Bayesian (NB) classifier and ID3 algorithm. A hybrid approach of combining Decision Tree (DT) and Support Vector Machine (SVM) was also proposed. [17] It defines about the ensemble approach which used Decision Tree (DT), Support Vector Machine (SVM) and hybrid DT-SVM classifier with waits. The ensemble approach resulted in 100 percent accuracy on the tested dataset. Various types of combinations are possible thus many approaches can be proposed and best resulting approaches can be implemented practically.

Combining supervised and unsupervised techniques:

There are number of unsupervised and supervised learning algorithms whose combinations can be made. [18] In the recent past years plenty such hybrid types are approached. By this the performance of supervised algorithm is eminently raised as accuracy of anomaly detection rate can be eminently enhanced by the help of unsupervised algorithms. Consolidation of k means and ID3 was prospected for classification of anomalous and normal activities in computer Address Resolution Protocol (ARP) traffic and accuracy of 98 percent.

5. COMPARISON

Methods used	Methodology	Pros	Cons
SVM classification and k-medoids clustering	Similar data occurrences are grouped by k- medoids type and resulting clusters are classified using SVM classifiers.	Higher accuracy.	Time complexity is [19] more when the dataset is very large.
k-Medoids Clustering and Naïve Bayes Classification	Similar data occurrences are grouped by using k-Medoids clustering technique. Resulting clusters are classified using Naïve Bayes classifiers.	Increase in detection Rate and reduction in mean time of false alarm rate.	Hard to predict when naïve Bayes classifier in the different environments.
One Class and Two Class Support Vector Machines (In cloud computing)	First class SVM is used for Revealing abnormality score. [20] Secondly detector is retrained when certain new data records are included in the existing dataset.	It does not require a prior failure history and is self-adaptive by learning from observed failure events.	The accuracy of failure detection cannot reach 100%.
Naive Bayes and decision tree for adaptive intrusion detection	It execute balance detections and keeps false positives at acceptable level for different types of network attacks.	Minimized false positives and maximized balance detection rates.	Require improvement of False positive rate to remote to user attacks.

Methods used	Methodology	Pros	Cons
k-Means clustering and ID3 decision tree learning Methods	k-Means clustering is first applied to the normal training instances to form k clusters. [21] An ID3 decision tree is constructed on each cluster.	Outperforms the individual k-Means and the ID3.	This approach is limited to specific Dataset.
Decision Tree (DT) and Support Vector Machines (SVM)	The data set is first developed through the DT and node information is achieved and is developed along with the original set of attributes through SVM to obtain the final output.	Delivers good performance on the KDD cup dataset.	This approach when correlated to SVM delivers equivalent results.
Ensemble approach	Information from different individual classifiers is combined to take the final decision.	Gave best performance for Probe and R2L classes. [22] 100% accuracy might be possible for other classes if proper base classifiers are selected.	Selection of base classifiers cannot be done automatically.

6. CONCLUSION

In this paper various data mining techniques are defined for the anomaly detection that had been proposed in the past few years. This review will be helpful for gaining a basic insight of various ways 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 other. Every day new unknown attacks are witnessed and thus there is a need of those approaches that can detect the unknown behavior in the data set stored, transferred or modified.

In this research work fusion or combination of already existing algorithms are mentioned that have been proposed. Interested ones on this field can combine the modified version of already existing algorithms. For example there are various new approaches in the modification of decision trees (such as ID3, C4.5), GA, SVM (including optimized and multiple kernel based approaches). This may yield more accurate results.

REFERENCES

[1]Chandola V, Banerjee A, Kumar V, “Anomaly detection: A survey”, ACM 2009, p.15.

[2] LeeW. Stolfo, J. Salvatore, “Data mining approaches for intrusion detection”, Proceedings of the 7th USENIX Security Symposium, San Antonio, Texas, 1998, p.9-94.

[3] Chauhan A., Mishra G. , Kumar G, “Survey on Data mining Techniques in Intrusion Detection”, International Journal of Scientific & Engineering Research, 2011, p.1-4.

[4] Padhy N, Mishra P , Panigrahi R., “The Survey of Data Mining techniques and Feature Scope”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2012, p. 43-58.

[5] Phua C, Lee V., Smith K., Gayler R., “A comprehensive survey of data mining-based fraud detection”, research, 2010, p. 1-14.

[6] Agarwal B, Mittal N, “Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques, Procedia Technology”, 6, 2012, p. 996-1003.

[7]A.M.Chandrashekhar and K. Raghuveer, “Confederation of FCM Clustering, ANN and SVM Techniques of Data mining to Implement Hybrid NIDS Using Corrected KDD Cup Dataset”, Communication and Signal Processing (ICCSP) IEEE International Conference,2014, Page 672-676.

[8] A.M.Chandrashekhar and K. Raghuveer , “Improvising Intrusion detection precision of ANN based NIDS by incorporating various data Normalization Technique – A Performance Appraisal”, IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014.

[9] A. M Chandrashekhar and K. Raghuveer, “Diverse and Conglomerate modi-operandi for anomaly intrusion detection systems”, International Journal of Computer Application (IJCA) Special Issue on “Network Security and Cryptography (NSC)”, 2011.

[10]A.M Chandrashekhar and K.Raghuveer,“Hard clustering vs. soft clustering: A close contest for attaining supremacy in hybrid NIDS Development”, Proceedings of International Conference on Communication and Computing (ICCC - 2014), Elsevier science and Technology Publications.

[11]A.M.Chandrashekhar, K.Raghuveer, “Amalgamation of K-means clustering algorithm with standard MLP and SVM based neural networks to implement network intrusion detection system”, Advanced Computing, Networking, and

Informatics –Volume 2(June 2014), Volume 28 of the series Smart Inovation, Systems and Technologies pp 273-283.

[12] A. M. Chandrashekhar and K. Raghuveer, "Fusion of multiple data mining techniques for effective network intrusion detection – A Contemporary approach", Proceedings of Fifth International Conference on Security of Information and Networks (SIN 2012), 2012, Page 178-182.

[13] A. M. Chandrashekhar, Jagadish Revapgol, Vinayaka Pattanashetti, "Big data security issues in networking", International Journal of Scientific Research in Science, Engineering and Technology (*IJSRSET),* Volume 2, Issue 1, JAN-2016.

[14] Puneeth L Sankadal, A. M Chandrashekhar, Prashanth Chillabatte, "Network Security situation awareness system" International Journal of Advanced Research in Information and Communication Engineering(IJARICE), Volume 3, Issue 5, May 2015.

[15]P.Koushik,A.M.Chandrashekhar,JagadeeshTakkalakaki, "Information security threats, awareness and cognizance" International Journal for Technical research in Engineering (IJTRE), Volume 2, Issue 9, May 2015.

[16] A.M.Chandrashekhar, Yadunandan Huded, H S Sachin Kumar, "Advances in Information security risk practices" International Journal of Advanced Research in data mining and Cloud computing (IJARDC), Volume 3, Issue 5, May 2015.

[17]A.M.Chandrashekhar,MukthaG,AnjanaD,"Cyberstalking and Cyberbullying: Effects and prevention measures", Imperial Journal of Interdisciplinary Research (IJIR), Volume 2, Issue 2, JAN-2016.

[18] A.M.Chandrashekhar, Syed Tahseen Ahmed, Rahul N, "Analysis of security threats to database storage systems" International Journal of Advanced Research in data mining and Cloud computing (IJARDC), Volume 3, Issue 5, May 2015.

[19]A.M.Chandrashekhar,K.K. Sowmyashree, RS Sheethal, "Pyramidal aggregation on communication security" International Journal of Advanced Research in Computer Science and Applications (IJARCSA), Volume 3, Issue 5, May 2015.

[20] A.M.Chandrashekhar, Rahil kumar Gupta, Shivaraj H. P, "Role of information security awareness in success of an organization" International Journal of Research(IJR), Volume 2, Issue 6, May 2015.

[21] A.M.Chandrashekhar, Huda Mirza Saifuddin, Spoorthi B.S, "Exploration of the ingredients of original security" International Journal of Advanced Research in Computer Science and Applications(IJARCSA), Volume 3, Issue 5, May 2015.

[22]A.M.Chandrasekhar, Ngaveni Bhavi, Pushpanjali M K, "Hierarchical Group Communication Security", International journal of Advanced research in Computer science and Applications (IJARCSA), Volume 4, Issue 1,Feb-2016.