Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 708 – 713

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Survey on Anomaly Detection using Data Mining Techniques

Shikha Agrawal, Jitendra Agrawal

Department of Computer Science and Engineering, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India

Abstract

In the present world huge amounts of data are stored and transferred from one location to another. The data when transferred or stored is primed exposed to 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 data mining techniques to detect the surprising behaviour hidden within data increasing the chances of being intruded or attacked. Various hybrid approaches have also been made in order to detect known and unknown attacks more accurately. 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.

Peer-review under responsibility of KES International

Keywords: Anomaly Detection, Clustering, Classification, Data Mining, Intrusion Detection System.

1. Introduction

Intrusion Detection Systems (IDS) are security tools that provided to strengthen the security of communication and information systems. This approach 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 or a hybrid detection system [29]. 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 behavior. Hybrid intrusion detection systems combine the techniques of both these approaches. Each technique has its own advantages and disadvantages. Few benefits of anomaly detection techniques over others can be stated as follows. Firstly, they are capable of detecting insider attacks. For example if any user is using any stolen account and perform such actions that are beyond normal profile of the user, an alarm will be generated by the anomaly detection system. Secondly, the detection system is based on custom made profiles. It becomes very difficult for an attacker to carry out any activity without setting off an alarm. Finally, it can detect the attacks that are previously not known. Anomaly detection systems look for anomalous events rather than the attacks. In this paper we focus upon the various anomaly detection techniques.

1.1. 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

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2015.08.220

be a surprising behavior which is previously not known. It may or may not be harmful. The anomaly detection provides very significant and critical information in various applications, for example Credit card thefts or identity thefts [1]. When data has to be analyzed in order to find relationship or to predict known or unknown data mining techniques are used. These include clustering, 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 approach the authors try to combine existing data mining algorithms to derive better results. Thus detecting the abnormal or unexpected behavior or anomalies will yield to study and categorize it into new type of attacks or any particular type of intrusions. This survey attempts to provide a better understanding among the various types of data mining approaches towards anomaly detection that has been made until now.

1.2. Basic Methodology of anomaly detection technique

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

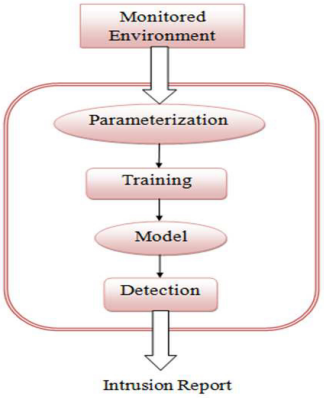

Figure 1: Methodology of Anomaly Detection

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 considered. 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.

2. 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 abnormal 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 main types and subtypes of known fraud, and also presented the nature of data evidence collected within affected industries [6]. Padhy et al (2012) provided a detailed survey of data mining applications and its feature scope. They stated that anomaly detection is an application of data mining where various data mining techniques can be applied [3] Amanpreet, Mishra, and Kumar (2012) described readymade data mining techniques that can be applied directly to detect the intrusion [7]. García et al (2009) have surveyed the most relevant works in the field of automatic network intrusion detection [15]. They provided a wide prospective to the techniques that they can be practically deployed by viewing the possible causes for the lack of acceptance to the proposed novel approaches.

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:

2.1. Clustering based Anomaly Detection techniques

Clustering can be defined as a division of data into group of similar objects. Each group, or cluster, consists of objects that are similar to one another and dissimilar to objects in other groups [13]. Clustering algorithms are able to detect intrusions without prior knowledge. There are various methods to perform clustering that can be applied for the anomaly detection. 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 grouped. 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 training 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 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 [11].
- *EM Clustering:* This algorithm can be viewed as an extension of k Means which assigns an object to the cluster to which it is similar, based on the mean of cluster. In this approach 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 boundaries in between the clusters. Here new mean is computed on the basis of weight measures [12]. When compared to k means and k medoids, EM outperformed them and resulted in higher accuracy [11].
- *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 [11]. It is based on the Nearest Neighbour algorithm and implements a well-defined distance metric to detect outliers. Greater the distance of the object to its neighbour, the more likely it is to be an outlier. It is an efficient approach in detecting probing attacks an Denial of Service (DoS) attacks. Other one is Density based local outlier approach. Distance based outlier detection depend on the overall or global distribution of the given set of data points. The data is not uniformly distributed thus the distance based approach encounter various difficulties during analysis of data. The main idea of this density based method is to assign to each data example a degree of being outlier, which is called the Local Outlier Factor (LOF). The outlier factor is local in the sense that only a restricted neighborhood of each object is considered [14]. 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 [8]. An approximated local outlier factor algorithm is also proposed which can be learned at the sink nodes for the detection model in WSN. These provide more efficient and accurate results.

2.2. 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 membership is known. The category can be termed as class label. Various instances can belong to one or many of the class labels. 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 attribute. 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, top-down 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].
- *Fuzzy Logic:* It is derived from fuzzy set theory which deals with reasoning that is approximate rather than precisely deduced from classical predicate logic. The application side of fuzzy set theory deals 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 various other fuzzy data mining techniques to extract patterns that represent normal behaviour for intrusion detection that describe a variety of modifications in the existing data mining algorithms in order to increase the efficiency and accuracy [17].

- *Naïve bayes network:* There are many cases where the statistical dependencies or the causal relationships between system variables exist. 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 Baysian 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. 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 [21]. When decision tree and baysian techniques are compared, though the accuracy of decision tree is far better but computational time of baysian network is low [19]. Hence, when the data set is very large it will be efficient to use NB models.
- *Genetic Algorithm:* It was introduced in the field of computational biology. These algorithms belong to the larger class of Evolutionary Algorithms (EA). They generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, selection, mutation and crossover. Since then, they have been applied 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 advantages of GA techniques reported in case of anomaly detection are high attack detection rate and lower false-positive rate [17].
- *Neural Networks:* It is a set of interconnected nodes designed to imitate the functioning of the human brain. Each node has a weighted connection to several other nodes in neighbouring layers. Individual nodes take the input received from connected nodes and use the weights together with a simple function to compute output values. Neural networks can be constructed for supervised or unsupervised learning [20]. The user specifies 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 variety of applications and producing more accurate results than other existing computational learning models. 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 based on one set of examples belonging to a particular class and no negative examples rather than using positive and negative example [18]. When compared to neural networks 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].

2.3. Hybrid approaches

Using any particular algorithm alone does not yield proper results. Now and then new attacks are registered thus using any single algorithm will not suffice. In past few years approaches have been made by either combining or merging different algorithms together.

- *Cascading supervised techniques:* Here various classification algorithms are merged together in order to obtain higher accuracy. A combination of naïve bayes and decision tree algorithm was proposed. This hybrid algorithm was tested in Knowledge Data Discovery (KDD) cup dataset and the accuracy achieved was 99 percent. It concentrated on the development of the performance of Naïve Bayesian (NB) classifier and ID3 algorithm [22]. A hybrid approach of merging Decision Tree (DT) and Support Vector Machine (SVM) was also proposed. It described 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 [28]. 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. In the recent past years many such hybrid methods are approached. By this the efficiency of supervised algorithm is highly increased as accuracy of anomaly detection rate can be highly improved by use of unsupervised algorithms. Combination of k means and ID3 was proposed for classification of anomalous and normal activities in computer Address Resolution Protocol (ARP) traffic and accuracy of 98 percent

was achieved [24]. A new approach for the detection of network attacks, which aims to study the effectiveness of the method based on machine learning in intrusion detection, including artificial neural networks and support vector machine was proposed. The experimental results obtained by applying this approach to the KDD CUP'99 data set demonstrate that the proposed approach performs high performance, especially to U2R and U2L type attacks [25]. A hybrid approach for combining entropy of network features and SVM have been proposed that outperformed individual entropy and SVM techniques [2]. Thus hybrid approaches yield better results as combining different techniques by overcoming the drawback of each other and resulting in higher accuracy of anomaly detection. Table1 presents few hybrid approaches proposed for anomaly detection:

Table 1 : Compilation of hybrid approaches for anomaly detection

Author Name	Methods used	Methodology	Pros and Cons
Chitrakar, Roshan, and Chuanhe (2012)	SVM classification and k-medoids clustering	Similar data instances are grouped by k- medoids technique and resulting clusters are classified into using SVM classifiers	Higher accuracy. Time complexity is more when the dataset is very large.
Chitrakar, Roshan, and Chuanhe (2012)	k-Medoids Clustering and Naïve Bayes Classification	Similar data instances 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 different environments.
Fu, Liu and Pannu(2012)	One Class and Two Class Support Vector Machines (In cloud computing)	First class SVM is used for detecting abnormality score. 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%.
Farid, Harbi, and Rahman (2010)	Naive bayes and decision tree for adaptive intrusion detection	It performs 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.
Yasami and Mozaffari (2009)	k-Means clustering and ID3 decision tree learning methods	k-Means clustering is first applied to the normal training instances to form k clusters. An ID3 decision tree is constructed on each cluster.	Outperforms the individual k-Means and the ID3. This approach is limited to specific dataset.
Peddabachigari, Abraham,Grosan and Thomas (2007)	Decision Tree (DT) and Support Vector Machines (SVM)	The data set is first passed through the DT and node information is generated and is passed 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 compared to SVM delivers equivalent results.
Peddabachigari, Abraham, Grosan and Thomas (2007)	Ensemble approach	Information from different individual classifiers is combined to take the final decision.	Gave best performance for Probe and R2L classes. 100% accuracy might be possible for other classes if proper base classifiers are selected. Selection of base classifiers cannot be done automatically.

3. Analysis and Recommendations

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 other. 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 stored, transferred or modified. In this research work fusion or combination of already existing algorithms are mentioned that have been proposed. Interested researchers 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 Computing Surveys (CSUR); 41(3); 2009;p. 15 .
2. Agarwal B., Mittal N., Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques, Procedia Technology; 6; 2012; p. 996-1003.
3. Padhy N., Mishra P. , Panigrahi R., The Survey of Data Mining Applications and Feature Scope; International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2(3) ;2012; p. 43-58.
4. Lee W., Stolfo J. Salvatore, Data mining approaches for intrusion detection; Proceedings of the 7th USENIX Security Symposium, San Antonio, Texas; 1998;p. 79-94.
5. Lee W., Stolfo S.J., Mok K.W., Adaptive intrusion detection: A data mining approach; Artificial Intelligence Review;14(6);2000; p. 533-567.
6. Phua C., Lee V., Smith K., Gayler R., A comprehensive survey of data mining-based fraud detection ; research; 2010; p. 1-14.
7. Chauhan A., Mishra G. , Kumar G. , Survey on Data mining Techniques in Intrusion Detection; International Journal of Scientific & Engineering Research ; 2(7), 2011; p.1-4.
8. Xu L., Yeh Y. R., Lee Y. J., Li J., A Hierarchical Framework Using Approximated Local Outlier Factor for Efficient Anomaly Detection; Procedia Computer Science ; 19; 2013; p. 1174-1181.
9. T. Pang-Ning, M. Steinbach, V. Kumar, Introduction to data mining, Library of Congress, 2006.
10. Munz,G., Li S., Carle G., Traffic Anomaly Detection Using K-Means Clustering; GI/ITG Workshop MMBnet; 2007;p.1-8.
11. 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.
12. Han J., Kamber M., Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.
13. Berkhin P., A survey of clustering data mining techniques;Grouping multidimensional data; Springer Berlin Heidelberg; 2006; p. 25-71.
14. 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
15. 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.
16. Wu S. Y., Yen E., Data mining-based intrusion detectors; Expert Systems with Applications; 36(3); 2009; p. 5605-5612.
17. 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.
18. Tang D. H., Cao Z.,Machine Learning-based Intrusion Detection Algorithm; Journal of Computational Information Systems;5(6); 2009; p. 1825-1831.
19. 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
20. 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.
21. 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.
22. 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.
23. Fu S., Liu J., Pannu H., A Hybrid Anomaly Detection Framework in Cloud Computing Using One-Class and Two-Class Support Vector Machines; In Advanced Data Mining and Applications; Springer Berlin Heidelberg; 2012; p. 726-738.
24. Yasami Y., Mozaffari S. P., A novel unsupervised classification approach for network anomaly detection by k-Means clustering and ID3 decision tree learning methods; The Journal of Supercomputing; 53(1); 2010; p. 231-245.
25. Tang D. H., Cao Z., Machine Learning-based Intrusion Detection Algorithms; Journal of Computational Information Systems; 5(6); 2009; p. 1825-1831.
26. Chitrakar R., Chuanhe H., Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes Classification, In Proceedings of 8th IEEE International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM); 2012; p 1-5.
27. Chitrakar R., . Chuanhe,H., Anomaly detection using Support Vector Machine classification with k-Medoids clustering; In Proceedings of IEEE Third Asian Himalayas International Conference on Internet (AH-ICI); 2012; p. 1-5.
28. Peddabachigari S., Abraham A., Grosan C., Thomas J., Modeling intrusion detection system using hybrid intelligent systems; Journal of network and computer applications; 30(1); 2007; p. 114-132.
29. Patcha A., Park J. M., An overview of anomaly detection techniques: Existing solutions and latest technological trends; Computer Networks; 51(12); 2007; p. 3448-3470.