An Overview of Anomaly Detection Techniques in Rapid miner

[1]Sherlin Suresh, [2]N. Aarthi, [3]S. Dhivya, [4]U. Surya
[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor, [4]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]Avinashilingam Institute for Home Science and Higher Education for Women-School of Engineering, Coimbatore, India

Abstract: The world has vast amounts of data that are stored and transferred from one location to a different. The information or data once transferred is exposed to attack. Though numerous techniques or applications are offered to shield the data, loopholes exist. Therefore to research data and have to work out numerous quite attack data mining techniques have emerged to make it less vulnerable. Anomaly detection uses these data techniques to detect the stunning behavior hidden within the data, increasing the probabilities of being intruded or attacked. Numerous hybrid approaches have additionally been created so as to sight noted and unknown attacks more accurately. This paper reviews numerous data mining techniques for anomaly detection to produce higher understanding among the present techniques which will facilitate interested researches to figure future in this direction.

***IndexTerms* - Intrusion detection, Stages of anomaly techniques in data mining, Anomaly detection algorithms, Classification of anomaly detection.**

I. INTRODUCTION

Intrusion detection systems are security tools that provided to strengthen the safety of communication and data systems. This approach is analogous to different measures like antivirus code, firewalls and access management schemes. These systems are classified as, nursing anomaly detection system or a hybrid detection system. Anomaly detection in data mining is the identification of things which do not change to associate expected pattern or different things in a data set. Hybrid intrusion detection systems mix the techniques of these approaches. Every technique has its own benefits and drawbacks. Few advantages of anomaly detection techniques over others explicit as follows. They are capable of detecting insider attacks. The detection system relies on custom created profile. Finally, it observes the attacks that are antecedently not famed. Anomaly detection systems rummage around for abnormal events instead of the attacks.

II ANOMALY DETECTION

Anomaly detection is the method of finding the patterns in every dataset whose behaviors are termed as anomalies or outliers. The anomalies cannot perpetually be categorized as an attack however it will be a shocking behavior that antecedently not known. The anomaly detection provides terribly important and demanding data in numerous applications for instance credit cards thefts. Once data needs to be analyzed so as to seek out relationship or to predict renowned or unknown data mining techniques are used. Hybrid approaches are being created so as to achieve higher level of accuracy in detection. Therefore detecting the unexpected behavior can yield to check categorize it into new style of attacks or any explicit type of intrusions. This paper makes an attempt to supply a far better understanding among the varied types of data mining approaches towards anomaly detection that has been created.

The non-conforming patterns are known as anomalies. Anomalies are patterns in data that don't adjust to a well outlined notion of traditional behavior. These anomalies are captivating to analyze. Unwanted noise within the data can also be found in there. Novelty detection aims at detecting antecedently unobserved patterns within the data. The challenges for anomaly detection are drawing the boundary between the normal and abnormal behavior, availability of labelled information, and noisy data. There are three types of anomalies namely point anomaly, contextual anomaly and collective anomaly.

III METHODOLOGY OF ANAMOLY DETECTION

3.1 Parameterization

Preprocess information in to a pre-established formats specified it's acceptable or in accordance with the targeted systems behavior.

3.2 Training stage

A model is constructed on the idea of traditional behavior of the system. There are alternative ways that may be opted reckoning on the kind of anomaly detection. It is each manual and automatic.

3.3 Detection stage

Once the model for the system is accessible, it is compared with the parameterized traffic. If the deviation found, exceeds from a pre outlined threshold then alarm will be triggered.

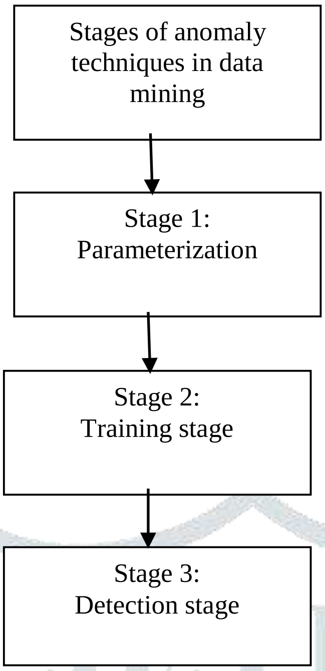
Stages of anomaly techniques in data mining
Stage 1: Parameterization
Stage 2: Training stage
Stage 3: Detection stage

Fig 3.1. Methodology of anomaly detection techniques

IV CLASSIFICATION BASED ANOMALY DETECTION

In case of anomaly detection it will classify the information into two classes specifically normal or abnormal. Following are common machine learning technologies in anomaly detection

4.1 Classification tree

In machine learning classification tree is also called as decision tree. It is a tree pattern graph that is similar to flow chart structure. The inner nodes are a test property every branch represents test result and last nodes represent the class. The most common algorithm used for classification tree is ID3 and C4.5. There are two ways for tree construction, top down tree construction and bottom up pruning.

4.2 Fuzzy logic

It is obtained from fuzzy set math that deals with reasoning that's approximately rather than precisely deduced from classical predicate logic. The appliance aspect of fuzzy set theory deals with well thought out globe skilled values for a complex downside. During this approach the information is classified on the idea of varied applied mathematics metrics. These parts of knowledge are applied with symbolic logic rules to classify them as normal or malicious. There are varied other fuzzy data processing techniques to extract patterns that represent normal behavior for intrusion detection that describe a spread of modifications within the existing data processing algorithms so as to extend the potency and accuracy.

4.3 Naive Bayes network

There are several cases wherever the applied math dependencies or the causal relationships between system variables exist. It will be troublesome to exactly specific the probabilistic relationships among these variables. In other words, the previous information concerning the system is simply that some variable can be influenced by others. To take advantage of this structural relationship between the random variables of a retardant, a probabilistic graph model called Naive Baysian Networks will be used. This model provides answer to the queries like if few discovered events are given then what is the chance of a specific kind of attack. It will or it may be done by mistreatment formula for conditional chance. The structure of a NB is usually delineated by a Directed Acyclic Graph. Where each node represents one among system variables every link encodes the influence of one node upon another. When decision tree and baysian techniques are compared, although the accuracy of decision tree is much higher but computational time of baysian network is low. Hence, when the information set is extremely massive it will be economical to use NB models.

4.4 Genetic algorithm

It has been introduced within the field of computational biology. These algorithms belong to the big category of organic process algorithms. They generate solutions to optimization problem practice techniques galvanized by natural evolution, like inheritance, selection, mutation and crossover. Since then, they have been applied in various fields with terribly promising results. In intrusion detection, the genetic algorithm is applied to derive a set of classification rules from the network audit information. The support-confidence framework is employed as a fitness function to evaluate the standard of every rule. The important properties of Genetic

Algorithm are its strength against noise and self-learning capabilities. The benefits of GA techniques according just in case of anomaly detection are high attack detection rate and lower false-positive rate.

4.5 Neural networks

It is a collection of interconnected nodes designed to imitate the functioning of the human brain. Each node features a weighted affiliation to many different nodes in near layers. Individual nodes take the input received from connected nodes and use the weights along with the calculated output values. Neural networks are often made for supervised or unattended learning. The user specifies the quantity of hidden layers further because the range of nodes among a selected hidden layer. Depending on the appliance the output layer of the neural network could contain one or many nodes. The Multilayer Perceptions neural networks are very productive in a form of applications and manufacturing a lot of correct results than different existing procedure learning models. They are capable of approximating to random accuracy, any continuous perform as long as they contain enough hidden units. This implies that such models will form any classification call boundary in feature space and therefore act as non-linear discriminate function.

4.6 Support Vector Machine

Support Vector Machine is wide applied to the field of pattern recognition. It is conjointly used for associate intrusion detection system. The one category SVM is predicated on one set of examples instead of mistreatment positive and negative example. In comparison to neural networks in KDD cup information set. It had been noticed that SVM out performed NN in terms of warning rate and accuracy in most quite attacks.

V RAPID MINER ANOMALY DETECTION EXTENSION

The Anomaly Detection Extension for Rapid Miner comprises the most unsupervised anomaly detection algorithms, assigning individual anomaly scores to data rows of data sets. It is used to find data, which is different from the normal, without the need for the data being labelled.

Anomaly detection algorithms could either be global or local. Global approaches refer to the techniques in which the anomaly score is assigned to each instance with respect to the entire dataset. The local approaches can detect outliers that are ignored using global approaches, especially in case of varying densities within a dataset. The anomaly detection extension contains two categories of approaches: nearest-neighbor based and clustering based algorithms. The first category assumes that outliers lie in sparse neighborhoods and that they are distant from their nearest neighbors. The second category operates on the output of clustering algorithms being thus much faster in general.

5.1 Nearest-Neighbor based algorithm

Nearest-neighbor based algorithms assign the anomaly score of data instances relative to their neighborhood. They assume that outliers are distant from their neighbors or that their neighborhood is sparse. The first assumption corresponds to k-NN which is a global anomaly detection approach, while the second assumption refers to local density based approaches. The k-NN global anomaly score is one of the most commonly used nearest-neighbor based algorithms. The anomaly score is either set to the average distance of the k-nearest-neighbors as proposed in or to the distance to the kth neighbor like the algorithm proposed in. Since the first method is much more robust to statistical fluctuations it should be preferred and it is also the method of choice in our experiments later on.

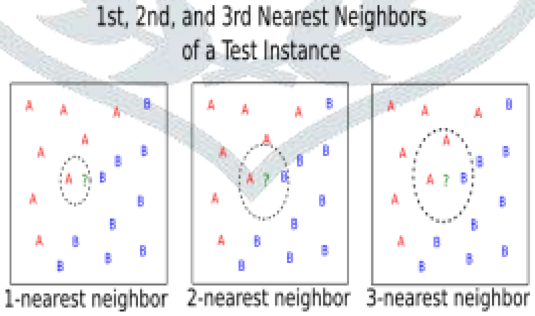

Fig 5.1. K-Nearest Neighbor

5.2 Clustering based algorithm

The process of arranging similar objects into groups is referred to as clustering. Clustering based anomaly detection techniques operate on the output of clustering algorithms, k-means algorithm. The anomalous instances either lie in sparse and small clusters, far from their cluster centroid or that they are not assigned to any cluster at all. The algorithms that implemented in extension use the output of any good clustering algorithm already available in Rapid Miner. The initial step followed by these algorithms is to classify the clusters into small and large clusters. Fig 3 shows clustering in which p is the distance to the cluster center of c1 is used for computing, c1 and c3 are identified as large clusters, while c2 is considered as small cluster, the white point illustrates the cluster centers.

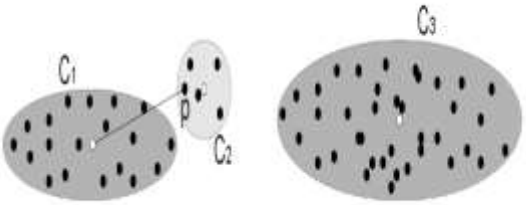

Fig 5.2. Clustering

VI CONCLUSION

In this paper numerous data mining techniques are represented for the anomaly detection that had been projected within the past few years. This review is going to be useful to researches for gaining basic insight of assorted approaches for the anomaly detection. Each day new unknown attacks are witnessed and so there is a necessity of these approaches that may sight the unknown behavior within the information set. For instance there are numerous new approaches within the modification call trees such as ID3, C4.5, GA, and SVM. This could yield additional correct results.

REFERENCES

[1] Survey on Anomaly Detection using Data Mining Techniques www.sciencedirect.com/science/article/pii/S1877050915023479

[2] Anomaly detection - Wikipedia https://en.wikipedia.org/wiki/Anomaly_detection

[3] Anomaly detection: A survey - UGC www.ugc.ac.in/mrp/paper/MRP-MAJOR-INFO-2013-17993-PAPER.pdf

[4] Mennatallah Amer, Markus Goldstein (2012), "Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for Rapid Miner"

[5] Shikha Agrawal, Jitendra Agrawal (2015), "Survey on Anomaly Detection using Data Mining Techniques"Procedia Computer Science 60, 708 – 713

[6] Amirah Mohamed Shahiria, WahidahHusaina, Nur'aini Abdul Rashid (2015), "A Review on Predicting Student's Performance using Data Mining Techniques", Procedia Computer Science 72 , 414 – 422

[7] http://rapidminerbook.com/index.php/chapter-downloads-13-24/chapter-23/

Ms.Sherlin Suresh B.E., Mtech.,
Assistant Professor,
Department of Computer Science and Engineering,
Avinashilingam Institute for Home Science and Higher Education for Women-School of Engineering,
Coimbatore.

Ms. N. Aarthi B.E.,M.E.,
Assistant Professor,
Department of Computer Science and Engineering,
Avinashilingam Institute for Home Science and Higher Education for Women-School of Engineering,
Coimbatore.

Ms. S. Dhivya B.E.,M.E.,
Assistant Professor,
Department of Computer Science and Engineering,
Avinashilingam Institute for Home Science and Higher Education for Women-School of Engineering,
Coimbatore.

Ms. U. Surya B.E.,M.E.,
Assistant Professor,
Department of Computer Science and Engineering,
Avinashilingam Institute for Home Science and Higher Education for Women-School of Engineering,
Coimbatore.