Available online at: http://www.ijmtst.com/vol6issue03.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 06, Issue No: 03, March 2020

A Review on Numerous Data Processing Ways for Outlier Detection

Shaik Khasim[1] | R Sravani[1] | Md Sirajul Huque[1]

[1]Department of Computer Science and Engineering, GNITC,Hyderabad, Telangana, India.

To Cite this Article

Shaik Khasim, R Sravani and Md Sirajul Huque, "A Review on Numerous Data Processing Ways for Outlier Detection", *International Journal for Modern Trends in Science and Technology*, Vol. 06, Issue 03, March 2020, pp.:58-64.

Article Info

Received on 10-February-2020, Revised on 19-February-2020, Accepted on 03-March-2020, Published on 08-March-2020.

ABSTRACT

Data repositories include huge quantity of knowledge being hold on at numerous locations and transferred from one location to different location. Once the information is moving or storing at some location it's exposed to attack. Different strategies or techniques offered to safeguard data from such attacks however still loopholes would possibly exist. Thus so as to safeguard knowledge from numerous attacks and create it less vulnerable numerous data processing techniques were used. Outliers detection uses these data processing strategies to spot the events or observations that raise suspicions of being intruded. Several hybrid approaches provided to spot well known as well as unknown data specifically. A review on numerous data processing approaches is provided during this paper to deal outlier's detection. We have additionally provided numerous outlier detection algorithms

KEYWORDS: *repositories, attacks, outliers, vulnerable, detection algorithms.*

I. INTRODUCTION

Intrusion Detection Systems (IDS) square measure security tools that provided to strengthen the safety of communication and knowledge systems. This approach is analogous to alternative measures like antivirus software system, firewalls and access management schemes. Conventionally, these systems are classified as a signature detection system, associate degree anomaly detection system or a hybrid detection system [29]. In signature primarily based detection, the system identifies patterns of traffic or application information is plausible to be malicious whereas anomaly detection systems compare activities against a standard outlined behavior. Hybrid intrusion detection systems mix the techniques of each these approaches. Every technique has its own benefits and drawbacks. Few advantages of anomaly detection techniques over others are often explicit as follows. Firstly, they're capable of police investigation corporate executive attacks. as an example if any user is victimization any taken account and perform such actions that square measure on the far side traditional profile of the user, an alarm are generated by the associate degree anomaly detection system. Secondly, the detection system relies on custom created profiles. It becomes terribly troublesome for associate degree wrongdoer to hold out associate degree activity while not setting off an alarm. Finally, it will find the attacks that square measure antecedently not identified. Anomaly detection

systems rummage around for abnormal events instead of the attacks. In this paper we tend to focus upon the varied anomaly detection techniques.

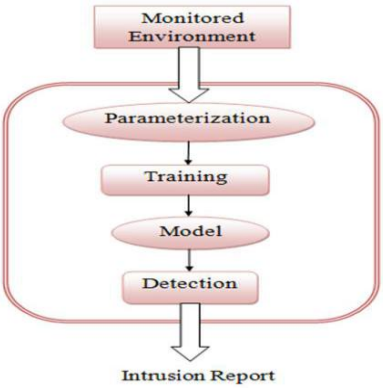
1.1. Anomaly Detection

Anomaly detection is that the method of finding the patterns during a dataset whose behavior is not traditional on expected. These surprising behaviors also are termed as anomalies or outliers. The Associate in nursing anomalies cannot perpetually be categorized as an attack however it are often a stunning behavior that is antecedently not far-famed. it's going to or might not be harmful. The anomaly detection provides terribly important and significant data in varied applications, as an example master card thefts or identity thefts [1]. Once information must be analyzed so as to seek out relationship or to predict far-famed or unknown data processing techniques square measure used. These embrace clump, classification and machine primarily based learning techniques. Hybrid approaches also are being created so as to achieve higher level of accuracy on police work anomalies. During this approach the authors try and mix existing data processing algorithms to derive higher results. Therefore police work the abnormal or surprising behavior or anomalies can yield to review and categorize it into new style of attacks or any explicit style of intrusions. This survey tries to produce a more robust understanding among the varied forms of data processing approaches towards anomaly detection that has been created up to now.

1.2. Basic Methodology of anomaly detection technique

Although totally different anomaly approaches exists, as shown in figure one parameter wise train a model before detection.

Parameterization: Pre process knowledge into a pre-established formats specified it's acceptable or in accordance with the targeted systems behavior.



Figure(1)

Training stage: A model is constructed on the idea of traditional (or abnormal) behavior of the system. There are a unit alternative ways which will be opted looking on the sort of anomaly detection thought-about. It will be each manual and automatic. Detection stage: once the model for the system is offered, it's compared with the (parameterized or the pre defined) determined traffic. If the deviation found exceeds (or is a smaller amount than once within the case of abnormality models) from a pre outlined threshold then an alarm are going to be triggered.

II. Anomaly Detection Mistreatment Data Processing Techniques

Anomalies are pattern within the knowledge that don't adapt to a well outlined traditional behavior. The explanation for anomaly could also be a malicious activity or some quite intrusion. This abnormal behavior found within the dataset is attention-grabbing to the analyst and this is often the most necessary feature for anomaly detection [14].Anomaly detection may be a topic that had been lined beneath numerous survey, review articles and books [4, 5]. Phua et al (2010) have done a close survey on numerous fraud detection techniques that has been dole out within the past few years. They need outlined the skilled fraudster, the most varieties and subtypes of far-famed fraud, and additionally given the character of knowledge proof collected within affected industries [6]. Padhy et al (2012) provided a close survey of knowledge mining applications and its feature scope. They explicit that associate degree anomaly detection is an application of knowledge mining wherever numerous data processing techniques is applied [3]Amanpreet, Mishra, and Kumar (2012) delineated readymade data processing techniques which will be applied on to sight the intrusion [7]. García et al (2009) have surveyed the foremost relevant works within the field of automatic network intrusion detection [15]. They provided a large prospective to the techniques that they will be much deployed by viewing the attainable causes for the lack of acceptance to the projected novel approaches. In this paper review of various approaches of anomaly detection focuses on the broad classification of existing data processing techniques. data processing consists of 4 categories of task; they're association rule learning, clustering, classification and regression. Next subdivision presents anomaly detection techniques beneath these four categories of task:

2.1 Clustering based Anomaly Detection techniques

Clustering is outlined as a division of information into cluster of comparable objects. Each group, or cluster, consists of objects that are similar to alternative and dissimilar to things in other teams [13]. Bunch algorithms are able to sight intrusions while not prior information. There are numerous strategies to perform bunch which will be applied for the anomaly detection. Following is that the description of a number of the planned approaches.

K-Means: k-Means bunch could be a cluster analysis methodology wherever we tend to outline k disjoint clusters on the premise of the feature price of the objects to be sorted. Here, k is that the user outlined parameter [9]. There has been a Network information Mining (NDM) approach that deploys the K-mean bunch algorithmic rule so as to separate time intervals with normal and abnormal traffic within the coaching dataset. The ensuing

- cluster centroids are then used for quick anomaly detection in observance of latest information [10].
- K-Medoids: This algorithmic rule is extremely like the k-Means algorithmic rule. It differs primarily in its illustration of the different clusters. Here every cluster is pictured by the foremost central object within the cluster, instead of by the implicit mean that will not belong to the cluster. The k-medoids methodology is additional sturdy than the k-means algorithmic rule within the presence of noise and outliers as a result of a medoid is a smaller amount influenced by outliers or different extreme values than a mean. This methodology detects network anomalies that contains unknown intrusion. it's been compared with numerous different clustering algorithms and are ascertain that once it involves accuracy, it produces far better results than k-Means [11].
- EM Clustering: This algorithmic rule is viewed as associate extension of k means that that assigns associate object to the cluster, to which it's similar, supported the mean of cluster. During this approach rather than assignment object within the dedicated cluster, assign the item to a cluster consistent with a weight representing the chance of membership. In different words there are no strict boundaries in between the clusters. Here new mean is computed on the premise of weight measures [12].When compared to k means that and k medoids, EM outperformed them and resulted in higher accuracy [11].

Outlier Detection Algorithms: Outlier detection could be a technique to search out patterns in information that don't adapt to expected behavior. Since associate outlier is outlined as a knowledge purpose that is extremely completely different from the remainder of the information, based on certain measures. There ar many outlier detection schemes. User will choose anybody of them on the premise of its efficiency and the way he will deal the matter of anomaly detection. One amongst the approach is Distance primarily based Approach [11]. It is supported the closest Neighbor algorithmic rule and implements a well-defined distance metric to sight outliers. Greater the space of the item to its neighbor, the additional seemingly it's to be associate outlier. It is an associate economical approach in detecting searching attacks associate Denial of Service (DoS) attacks. Different one is Density primarily based native outlier approach. Distance primarily based outlier detection rely upon the general or world distribution of the given set of information points. The data is not uniformly distributed so the space primarily based approach encounter numerous difficulties throughout

- analysis of information. The main plan of this density primarily based methodology is to assign to every information example a degree of being outlier, which is called the native Outlier issue (LOF). The outlier issue is native within the sense that solely a restricted neighborhood of each object is taken into account [14]. Numerous different algorithms ar planned for anomaly detection within the Wireless device Networks (WSN). A hierarchical framework are planned to beat challenges in WSN's wherever associate accurate model and therefore the approximated model is formed learned at the remote server and sink nodes [8]. An approximated native outlier issue algorithmic rule is additionally planned which might be learned at the sink nodes for the detection

model in WSN. These offer additional economical and correct results.

2.2 Classification of primarily based anomaly detection

Classification is outlined as a tangle of characteristic the class of latest instances on the premise of a coaching set of information containing observations (or instances or tuples) whose class membership is understood. The class is termed as category label. Various instances will belong to at least one or several of the category labels. In machine learning, classification is taken into account as associate instance of supervised learning for instance learning wherever a coaching set of correctly-identified observations is accessible. Associate algorithmic rule that implements classification is understood as a classifier. It is made to predict categorical labels or category label attribute. In case of anomaly detection it'll classify the information typically into 2 classes particularly traditional or abnormal. Following are common machine learning technologies in anomaly detection.

- Classification Tree: In machine learning classification tree is additionally referred to as a prediction model or call tree. It is a tree pattern graph that is analogous to flow chart structure; the inner nodes are a take a look at property, every branch represents take a look at result, and final nodes or leaves represent the category to that any object belongs. The foremost basic and common algorithmic rule used for classification tree is ID3 and C4.5 There are 2 strategies 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 compared to naïve Thomas Bayes classification, the result obtained from call trees was found to be correct [19].
- Fuzzy Logic: it's derived from fuzzy pure mathematics that deals with reasoning that's approximate instead of exactly deduced from classical predicate logic. The applying aspect of fuzzy pure mathematics deals with well thought out globe expert values for a posh drawback. During this approach the information is classedon the premise of assorted applied mathematics metrics.
- These parts of knowledge area unit applied with symbolic logic rules to classify them as traditional or malicious. There are a unit varied other fuzzy data processing techniques to extract patterns that represent traditional behavior for intrusion detection that describe a range of modifications within the existing data processing algorithms so as to extend the potency and accuracy [17].
- Naïve mathematician network: There are several cases wherever the applied math dependencies or the causative relationships between system variables exist. It is tough to exactly categorical the probabilistic relationships among these variables. In other words, the previous data regarding the system is solely that some variable may well be influenced by others. To take advantage of this structural relationship between the random variables of a tangle, a probabilistic graph model called Naïve Baysian Networks (NB) is used. This model provides answer to the queries like if few determined events area unit given then what the likelihood of a specific quite is attack is. It is done by mistreatment formula for conditional likelihood. The structure of a NB is usually depicted by a Directed Acyclic Graph (DAG) wherever each node represents one in every of system variables link encodes the influence of 1 node upon another [21]. When call tree and Bayesian techniques area unit compared, though the accuracy of call tree is much higher however computational time of Bayesian network is low [19]. Hence, once the information set is extremely giant it will be economical to use NB models.
- Genetic Algorithm: it absolutely was introduced within the field of procedure biology. These algorithms belong to the larger class of organic process Algorithms (EA). They generate solutions to improvement issues mistreatment techniques impressed by natural evolution, like inheritance, selection, mutation and crossover. Since then, they need been applied in various fields with terribly promising results. In intrusion detection, the Genetic rule (GA) is applied to derive a set of classification rules from the network audit knowledge. The support-confidence framework is used as a fitness function to evaluate the standard of every rule. Important properties of GA area

unit its lustiness against noise and self learning capabilities. The benefits of GA techniques reportable just in case of anomaly detection area unit high attack detection rate and lower false-positive rate [17].

- Neural Networks: it's a collection of interconnected nodes designed to imitate the functioning of the human brain. Each node incorporates a weighted affiliation to many alternative nodes in neighbor layers. Individual nodes take the input received from connected nodes and use the weights in conjunction with a straightforward operate to reckon output values. Neural networks are made for supervised or unattended learning [20]. The user specifies the quantity of hidden layers additionally because the variety of nodes among a selected hidden layer. reckoning on the applying, the output layer of the neural network might contain one or many nodes. The Multilayer Perceptions (MLP) neural networks are very successful form of applications and manufacturing a lot of correct results than alternative existing procedure learning models. They are capable of approximating to random accuracy, any continuous operate as long as they contain enough hidden units. This implies that such models will kind any classification call boundary in feature space and so act as non-linear discriminate operate.
- Support Vector Machine: This is a supervised learning approach used for classification and regression. Support Vector Machine (SVM) is wide applied to the sphere of pattern recognition. It is conjointly used for associate degree intrusion detection system. The one category SVM is predicated on one set of examples happiness to a specific category and no negative examples instead of mistreatment positive and negative example [18]. In comparison to neural networks in KDD
  cup knowledge set, it absolutely was identified that SVM out performed NN in terms of warning rate and accuracy in most quite attacks [18].

2.3. Hybrid approaches

Using any explicit rule alone doesn't yield correct results. Currently so new attacks area unit registered so mistreatment any single algorithm won't live up to. In past few years approaches are created by either combining or merging totally different algorithms together.

- Cascading supervised techniques: Here varied classification algorithms area unit incorporate along so as to get higher accuracy. a mixture of naïve mathematician and call tree rule was projected. This hybrid rule was tested in data knowledge Discovery (KDD) cup dataset and therefore the accuracy achieved was 99%. It targeted on the event of the performance of Naïve Bayesian (NB) classifier and ID3 rule [22]. A hybrid approach of merging call Tree (DT) and Support Vector Machine (SVM) was conjointly projected. It delineated regarding the ensemble approach that used call Tree (DT), Support Vector Machine (SVM) and hybrid DT-SVM classifier with waits. The ensemble approach resulted in one hundred pc accuracy on the tested dataset [28]. Varied styles of combinations area unit attainable so several approaches is projected and best ensuing approaches is enforced practically.
- Combining supervised and unattended techniques: There are a unit variety of unattended and supervised learning algorithms whose combos is created. Within the recent past years several such hybrid ways area unit approached. By this the potency of supervised rule extremely magnified as accuracy of anomaly detection rate is highly improved by use of unattended algorithms. Combination of k means that and ID3 was projected for classification of anomalous and traditional activities in computer code Resolution Protocol (ARP) traffic and accuracy of ninety eight percent was achieved [24]. a brand new approach for the detection of network attacks, that aims to review the effectiveness of the method supported machine learning in intrusion detection, together with artificial neural networks and support vector machine was planned. The experimental results obtained by applying this approach to the KDD CUP'99 knowledge set demonstrate that the planned approach performs high performance, particularly to U2R and U2L sort attacks [25]. A hybrid approach for combining entropy of network options and SVM are planned that outperformed individual

entropy and SVM techniques [2]. So hybrid approaches yield higher results as combining completely different techniques by overcoming the downside of every different and leading to higher accuracy of anomaly detection. Table1 presents few hybrid approaches planned for anomaly detection

Table 1: Comparison 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 Data set 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.

III. Analysis and Proposals

In this paper varied data processing techniques square measure delineated for the anomaly detection that had been projected within the past few years. This review are useful to researchers for gaining a basic insight of varied approaches for the anomaly detection. Although much work had been done victimization freelance algorithms, hybrid approaches square measure being immensely used as they supply higher results and overcome the downside of 1 approach over the opposite. on a daily basis new unknown attacks square measure witnessed and therefore there's a

need of these approaches which will observe the unknown behavior within the knowledge set keep, transferred or changed. During this analysis work fusion or combination of already existing algorithms square measure mentioned that are projected. Interested researchers will combine the changed version of already existing algorithms. As an example there square measure varied new approaches within the modification of call trees (such as ID3, C4.5), GA, SVM (including optimized and multiple kernel based mostly approaches). This could yield more correct 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 victimization data processing Techniques, Procedia Technology; 6; 2012; p. 996- 1003.

[3] Padhy N., Mishra P. , Panigrahi R., The Survey of information Mining Applications and have Scope; International Journal of technology, Engineering and data Technology (IJCSEIT), 2(3) ;2012;p.43-58.

[4] Lee W., Stolfo J. Salvatore, data processing approaches for intrusion detection; Proceedings of the seventh USENIX Security conference, urban center, Texas;1998;p. 79-94.

[5] Lee W., Stolfo S.J., Mok K.W., adaptative intrusion detection: a knowledge mining approach; AIReview;14(6);2000;p.533-567.

[6] Phua C., Lee V., Smith K., Gayler R., A comprehensive survey of information mining-based frauddetection;research;2010;p.1-14.

[7] Chauhan A., Mishra G. , Kumar G. , Survey on data processing Techniques in Intrusion Detection; International Journal of Scientific & Engineering analysis ; 2(7), 2011; p.1-4.

[8] Xu L., Yeh Y. R., Lee Y. J., Li J., A hierarchical Framework victimization Approximated native Outlier issue for economical Anomaly Detection; Procedia technology ; 19; 2013; p. 1174-1181.

[9] T. Pang-Ning, M. Steinbach, V. Kumar, Introduction to data processing, Library of Congress, 2006.

[10] Munz,G., Li S., Carle G., Traffic Anomaly Detection victimization K-Means Clustering; GI/ITGWorkshopMMBnet;2007;p.1-8.

[11] Syarif I., Prugel-Bennett A., Wills G., data processing approaches for network intrusion detection from spatiality reduction to misuse and anomaly detection; Journal of knowledge TechnologyReview;3(2);2012;p.70-83.

[12] Han J., Kamber M., information Mining: ideas and Techniques, second edition, Morgan Kaufmann,2006.

[13] Berkhin P., A survey of bunch data processing techniques;Grouping dimensional data; SpringerBerlinHeidelberg;2006;p.25-71.

[14] Dokas P., . Ertoz L., Kumar V., Lazarevic A., Srivastava J., Tan P. N., data processing for network intrusion detection, In Proceedings of independent agency Workshop on Next GenerationinformationMining;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., information mining-based intrusion detectors; skilled Systems with Applications;36(3);2009;p.5605-5612.

[17] Kaur N., Survey paper on data processing 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 procedureinfoSystems;5(6);2009;p.1825-1831.

[19] Amor N. B., Benferhat S., Elouedi Z., Naive mathematician vs call trees in intrusion detection systems, In Proceedings of the ACM conference 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; skilled Systems with Applications; 36(10); 2009; p. 11994-12000.

[22] Farid D. M., Harbi N., Rahman M. Z., Combining naive mathematician and call tree for adaptative 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 victimization One-Class and Two-Class Support Vector Machines; In Advanced data processing and Applications; Springer BerlinHeidelberg;2012;p.726-738.

[24] Yasami Y., Mozaffari S. P., a completely unique unsupervised classification approach for network anomaly detection by k-Means bunch and ID3 call tree learning methods; The Journal ofSupercomputing;53(1);2010;p.231-245.

[25] Tang D. H., Cao Z., Machine Learning-based Intrusion Detection Algorithms; Journal of procedureinfoSystems;5(6);2009;p.1825-1831.

[26] Chitrakar R., Chuanhe H., Anomaly primarily based Intrusion Detection victimization Hybrid Learning Approach of mixing k-Medoids bunch and Naïve mathematician Classification, In Proceedings of eighth IEEE International Conference on Wireless Communications, NetworkingandMobileComputing(WiCOM);2012;p1-5.

[27] Chitrakar R., . Chuanhe,H., Anomaly detection victimization Support Vector Machine classification with k-Medoids clustering; In Proceedings of IEEE Third Asian mountain chain International Conference on web (AH-ICI); 2012; p. 1-5.

[28] Peddabachigari S., Abraham A., Grosan C., Thomas J., Modeling intrusion detection system victimization hybrid intelligent systems; Journal of network and pc applications; 30(1); 2007; p. 114-132.

[29] Patcha A., Park J. M., an summary of anomaly detection techniques: Existing solutions and latest technological trends; pc Networks; 51(12); 2007; p. 3448-3470