International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 02 | Feb 2021 www.irjet.net p-ISSN: 2395-0072

Survey on Anomaly Detection

Vanishree A[1], Sumathi L[2]

[1]*M.E. Student,Department of Computer Science and Engineering,Government College of Technology, Coimbatore, India.*

[2]*Assistant Professor, Department of Computer Science and Engineering,Government College of Technology, Coimbatore, India.*

Abstract - *Identifying abnormal behaviour in network traffic at an early stage is vital to avoid severe disaster. Researchers explored different ways to detect abnormal behaviour. In an environment this paper focuses on techniques that learn normal behaviour of network traffic patterns to detect abnormal. Survey reviews anomaly detection methods to lay out better understanding among existing state of art methods that may help implicated researchers to work further in the direction.*

Key Words: **Anomaly Detection, Intrusion Detection, Classification, Clustering, Statistical method.**

1.INTRODUCTION

Intrusion detection system plays an important role in research and development with an increase in attacks on computers and networks. An intrusion detection system monitors the events occurring in a computer system or networks for analysing the patterns of intrusions. There are two basic intrusion detection systems (IDS) approaches: misuse detection (signature-based) and anomaly detection.

1.1 Misuse detection method:

The misuse detection method uses well-known recognized patterns to match and identify the intrusions. It performs pattern matching between the captured attack signatures and network traffic. If a match is detected, the system produces an alarm. The main advantage of the signature detection model is that it can accurately detect instances of known attacks. The main disadvantage is that it lacks the ability to detect new anomalies or zero-day attacks [10].

1.2 Anomaly detection method:

Anomaly-based network intrusion detection systems are presently a principal focus of researchers in the field of intrusion detection. An anomaly detection method monitors the behaviour of a system and flags significant deviations from the normal activity as anomalies [9].

Anomaly detection is about finding unusual event or pattern anomalies in historical data using mathematical methods. In data analysis, there are two ways (outlier detection and novelty detection) to search for anomalies. The outlier is the data information that differs from other data points present in the train dataset. The novelty point also varies from other information in the dataset, but unlike outliers, novelty points emerge in the test dataset and usually absent in the train dataset [8]. The most common reason for the outliers are:

- Errors in data information.
- Noise in data points.
- Hidden patterns in the dataset.

Noise data points should be removed and the errors in data should be rectified. The detection of anomalous things can be useful in fraud detection and intrusion detection. The main goal of Anomaly Detection analysis is to find the unusual pattern in the data observations.

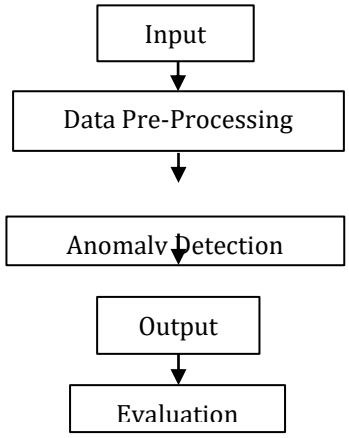

Fig -1: Overall Framework for Anomaly Detection

Fig.1. displays an overall framework for anomaly detection. The input data requires pre-processing in order to eliminate the unwanted attributes. Then anomaly detection techniques are employed to find the anomalies in the historical data by fitting the historical data. The main advantage of anomaly detection is, it does not require prior knowledge of an intrusion and thus can detect new anomalies. The main disadvantage is, it should not be ready to describe what constitutes an attack and should have a high false-positive rate [10]. One important issue for anomaly detection is how anomalies are represented as output which, generally, is in one of the two following ways (Chandola et al., 2009).

SCORES: Scoring-based anomaly detection methods assign an anomaly score to each data object. Then, the scores are ranked and an analyst chooses the outliers and the corresponding anomaly scores within a range from 0 to 1.

BINARY/LABEL: According to these techniques, outputs are considered in a binary fashion, i.e., either anomalous or normal. Techniques which provide binary labels are computationally efficient since each data object does not need to provide or have an anomaly score.

The rest of the paper is organized as follows: Section 2 briefly discusses basic anomaly detection techniques. Section 3 discusses experimental results of sample anomaly detection algorithms. Section 4 concludes the paper.

2. ANOMALY DETECTION TECHNIQUES

Varun Chandola et al. [18] have done a detailed survey on various anomaly detection techniques that have been carried out in the past few years. In this paper, three approaches in anomaly detection were discussed as clustering, classification, and statistical techniques.

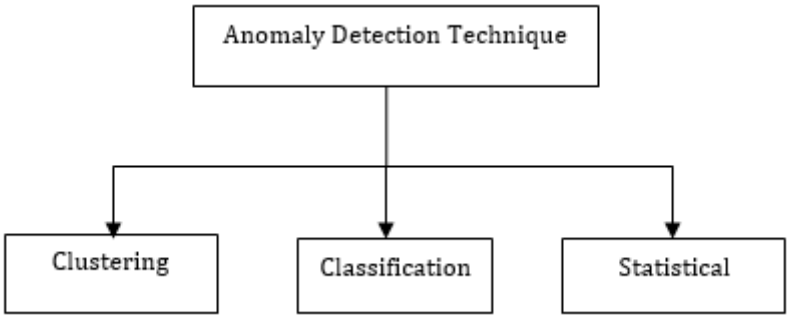

Fig -2: Techniques of Anomaly Detection

2.1 Clustering based Anomaly Detection Techniques

Clustering is the most common exploratory data analysis method. The cluster analysis can be done on the basis of attributes to identify subgroups of attributes based on the training sample. It can be defined as a division of data into groups of similar objects.

2.1.1 K-Means as Anomaly Detection [10]:

K-Means is the clustering algorithm used to identify the group of similar objects in training data. It divides the group of objects into a pre-defined number of clusters specified by the user. The first initial step is to select a set of k instances as centers of the clusters. Here k is a predefined parameter [7]. Next, the algorithm selects each data instance from the data set and assigns it to the closest cluster. There are different ways to identify the distance between data points and the centroid and Euclidean distance is the widely used method in the literature. The cluster centers are always recalculated based on the average of each cluster data point. This process is iterated until no more changes are made.

2.1.2 K-Medoids as Anomaly Detection [12]:

K-Medoids is a partition-based clustering algorithm which is similar to the K-Means algorithm. In this method, the most centrally located point is considered as medoid (i.e. centroid) rather than taking the mean value as a centroid. Based on their distance partitioning can be performed. The k-medoids is more robust than the K-Means algorithm in the existence of noise and outliers because a medoid is less controlled by outliers or other excessive values than a mean.

2.1.3 Agglomerative clustering as Anomaly Detection [29]:

Mazarbhuiya et al. [3] proposed an agglomerative hierarchical clustering technique for anomaly detection. Initially, define the similarity measure of two data points as a weighted collection of their numeric features and their categorical features. In [29][26], the Canberra metrics are used as a distance measure to find out the similarities between the numerical features. Agglomerative type of clustering produces better performance in finding the anomalies in the network dataset.

2.1.4 DBSCAN as Anomaly Detection [13]:

Tran Manh Thang et al. [13] presented a new way to use Density-based clustering methods with multiple parameters. DBSCAN depends on two parameters, namely epsilon and minpts. The Epsilon parameter represents how similar two network connections are and the minpts represent how many network connections that have similar characteristics exist within epsilon distance. To identify the minpts parameter for each cluster, the number of neighbour points is identified. By applying for this extension, DBSCAN can accurately identify the anomalies and different types of network traffic.

2.2 Classification based Anomaly Detection Techniques

Classification can be defined as a problem of identifying the group of new instances on the basis of a training set of data containing instances whose group membership is known. The group can be termed as a class label. Anomaly detection will classify the information generally into two groups namely normal or abnormal. Following are the most common anomaly detection techniques:

2.2.1 Decision tree as Anomaly Detection [28]:

In machine learning, the decision tree is also called a prediction method or classification tree. It is a tree pattern graph which is similar to flow chart form; the internal nodes are a test property, each branch denotes test result, and final nodes or leaves denote the class to which any object belongs. The most common algorithm used for classification trees is ID3 and C4.5. There are two models for tree construction, top-down tree construction, and bottom-up pruning. ID3 and C4.5 belong to the top-down tree construction method. Further decision tree methods when compared to the Naive Bayes method, the result obtained from decision trees was found to be more accurate.

2.2.2 Random Forest as Anomaly Detection [13]:

Random Forest is an ensemble method, which predicts based on the results of a collection of Decision Trees. Each decision tree within the forests is built with a different bootstrap sample drawn from the original data set. Each tree is then constructed to the maximum size without any pruning [2].

Lee et al. [15] aimed to create a method to quantify the intensity of anomalous behaviour in network traffic and used Random Forest modelling for this purpose. When feature selection and parameter optimization, they proposed a Random Forest model employing a training dataset and on the detection section, calculate proximities between the evaluated flow and the normal data. Two ways in which to quantify the anomalous intensity of a flow: (i) the use of the lowest and highest proximities and the training normal flows; (ii) the employ of the proximity and the cluster centre of the testing normal data found by applying a clustering algorithm. The abnormal thresholds are computed averaging these values over a validation dataset. In this approach, the Random Forest model is employed primarily to calculate proximities.

2.2.3 Naive Bayes network as Anomaly Detection [11]:

In several cases, causal relationships between system variables exist. It can be challenging to precisely express the probabilistic relationships among these variables. To take advantage of this structural relationship between the random variables of a problem, a probabilistic graph model is used by Naive Bayes networks. In anomaly detection, this model provides answers to the queries like if few determined events are given then what's the probability of a specific reasonable attack. When decision tree, random forest and bayesian methods are compared, though the accuracy of the decision tree, random forest is far better but the performance score of Naive Bayes network is low [1], [15]. Hence, when the data set is very large it will be efficient to use Naive Bayes models.

2.2.4 Support Vector Machine as Anomaly Detection [19]:

The Support Vector Machine (SVM) method is to solve a hyperplane that maximizes the separating margin between the positive and negative classes. A remarkable property of SVM is its associate implementation of the structure risk minimization principle, based on statistical learning theory. The quality SVM method is a supervised learning method, which requires labelled data to create a classification rule. However, it can also be adapted as an unsupervised learning method whereby it tries to separate the entire set of training data from its origin whereas the regular supervised SVM method attempts to separate two classes of data in feature space by a hyperplane.

2.2.5 One Class Support Vector Machine as Anomaly Detection [24]:

Eskin et al. [24] analyzed the unsupervised SVM method employed to detect anomalous events. The algorithm finds hyper planes that separate the information instances from their origins with the maximal margin and then an optimization problem is solved to determine the best hyperplane. Employing a similar idea to it of the One-class SVM (OCSVM) method however during a supervised manner, a new method called Registry Anomaly Detection is developed to monitor Windows registry queries. It's usual that in normal computer activity, a definite set of registry keys is accessed by the Windows program. Based on the fact that users tend to frequently employ certain programs and registry activities are normal, deviations from these activities will be considered abnormal. The OCSVM is applied to the Registry Anomaly Detection system to detect abnormal data in the Windows registry. Registry Anomaly Detection maps the input information into a high-dimensional feature space via a kernel and iteratively finds the highest margin hyperplane to separate two categories of information.

Hu et al.[21] presented an anomaly detection technique that removes noisy information and is developed employing the Robust SVM (RSVM). In follow, training data often contain noise that invalidates the most assumption of the SVM that each one of the sample data for training is independently and identically distributed. As a result, the standard SVM results in an extremely non-linear decision boundary that results in poor generalization. During this situation, the Robust Support Vector Machine incorporates the averaging method in the form of a class origin to make the decision surface smoother and automatically control regularization. In addition, the number of support vectors within the RSVM is significantly less than the standard Support Vector Machine which results in a reduced run time.

2.3 Statistical based Anomaly Detection Techniques

Intrusion detection techniques have also been developed using statistical theories; for example, the established chi-square theory is used for anomaly detection [25]. According to this method, a profile of normal class in a data system is created. The basic idea in this method is to detect both a large departure of classes from normal as anomalous and intrusions.

2.3.1 Linear Discriminant Analysis as Anomaly Detection [6]:

Linear Discriminant Analysis is a well-known statistical method used as a dimensionality reduction technique. Subba et al. [6] presented a Linear Discriminant Analysis method denotes the n-dimensional dataset into a smaller k-dimensional dataset (k< n), while at the same time maintains all the relevant class discrimination information. In this paper, do not use the LDA as a feature reduction technique; also analyze how it can be used to build an anomaly detection method.

2.3.2 Principal component analysis (PCA) as Anomaly Detection [18]:

Shyu et al. [18] presented an easier way to analyze high dimensional network traffic dataset using Principal component analysis. Principal component analyses are linear combinations of p random variables (A1; A2; ...Ap) and can be characterized are uncorrelated, with their variances sorted in order from maximum to minimum, their total difference equal to the difference of the original data.

An anomaly detection technique based on principal component analysis [18] has the advantages of:

- being free from any assumption of applied statistical distribution;
- being ready to minimize the dimension of the data while not losing any vital information;
- having negligible process complexness that supports real-time anomaly detection.

2.3.3 Signal processing technique as Anomaly Detection [23]:

The signal processing method is a remarkable analysis space, employing a method for anomaly detection has hardly been explored. Thottan and Ji et al. [23] a statistical signal processing technique based on abrupt change detection is presented. This method describes network anomalies in two ways:

- Anomalies correspond to network failures and performance problems;
- encompasses security-related issues such as DoS attacks.

Thottan and Ji et al. [23] management information bases are used to produce a network health function, which is employed to raise alarms corresponding to anomalous networks. The unusual behaviours in these bases are determined by finding rapid changes in their statistics. A hypothesis test based on the general likelihood ratio (GLR) is used to detect the changes to provide the range of abnormality on a scale of between 0 and 1.

3. RESULTS AND DISCUSSION

A technique for finding unusual events termed as anomaly detection. It has been experimented on the UNSW-NB15 dataset. It has been split into 60% for training and 40% for testing. The comparison among classification, clustering and statistical methods has been done. One-Class SVM is one of the best classification method helps in anomaly detection, which has been applied in our dataset gives 83.4% (accuracy), 63% (sensitivity) of anomalies were identified, 100% (specificity) of its detection were correct, whereas k-means for clustering method gives 64.8%(accuracy), 42.8% (sensitivity) of anomalies were identified,27.8% (specificity) of its detection were correct and principal component analysis for statistical method gives 44.9% (accuracy), 100% (sensitivity)) of anomalies were identified, 32% (specificity) of its detection were correct.

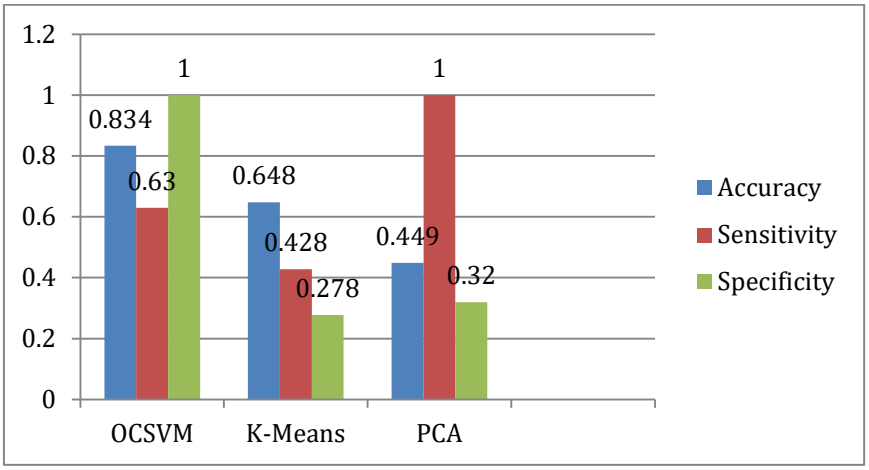

Chart -1: Comparison of performance metrics

4. CONCLUSIONS

The paper surveys some of the techniques under basic three analysis of anomaly detection. Experimental results on UNSW-NB15 dataset using one class SVM, K-means and PCA based anomaly detection is discussed which can be used as a starting point for the naïve research community in the area of network traffic anomaly detection analysis.

REFERENCES

[1] Estévez-Pereira, J.J., Fernández, D. and Novoa, F.J., 2020. Network Anomaly Detection Using Machine Learning Techniques. In Multidisciplinary Digital Publishing Institute Proceedings (Vol. 54, No. 1, p. 8).

[2] Resende, P.A.A. and Drummond, A.C., 2018. A survey of random forest based methods for intrusion detection systems. ACM Computing Surveys (CSUR), 51(3), pp.1-36.

[3] Mazarbhuiya, F.A., AlZahrani, M.Y. and Georgieva, L., 2018, June. Anomaly Detection Using Agglomerative Hierarchical Clustering Algorithm. In International Conference on Information Science and Applications (pp. 475-484). Springer, Singapore.

[4] Kumari, R., Singh, M.K., Jha, R. and Singh, N.K., 2016, March. Anomaly detection in network traffic using K-mean clustering. In 2016 3rd International Conference on Recent Advances in Information Technology (RAIT) (pp. 387-393). IEEE.

[5] Vasan, K.K. and Surendiran, B., 2016. Dimensionality reduction using Principal Component Analysis for network intrusion detection. Perspectives in Science, 8, pp.510-512.

[6] Subba, B., Biswas, S. and Karmakar, S., 2015, December. Intrusion detection systems using linear discriminant analysis and logistic regression. In 2015 Annual IEEE India Conference (INDICON) (pp. 1-6). IEEE.

[7] Tripathy, S. and Sahoo, P.L., 2015. A survery of different methods of clustering for anomaly detection. International Journal of Science and Engineering Research, 6(1).

[8] Zhang, M.; Xu, B. & Gong, J. (2015), An Anomaly Detection Model Based on One-Class SVM to Detect Network Intrusions., in 'MSN' , IEEE Computer Society, , pp. 102-107.

[9] Muniyandi, A.P., Rajeswari, R. and Rajaram, R., 2012. Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm. Procedia Engineering, 30, pp.174-182

[10] Syarif, I., Prugel-Bennett, A. and Wills, G., 2012, April. Unsupervised clustering approach for network anomaly detection. In International conference on networked digital technologies (pp. 135-145). Springer, Berlin, Heidelberg.

[11] Mukherjee, S. and Sharma, N., 2012. Intrusion detection using naive Bayes classifier with feature reduction. Procedia Technology, 4, pp.119-128.

[12] Chitrakar, R. and Chuanhe, H., 2012, November. Anomaly detection using support vector machine classification with k-medoids clustering. In 2012 Third Asian Himalayas International Conference on Internet (pp. 1-5). IEEE.

[13] Thang, T.M. and Kim, J., 2011, April. The anomaly detection by using dbscan clustering with multiple parameters. In 2011 International Conference on Information Science and Applications (pp. 1-5). IEEE.

[14] Chandola, V., Banerjee, A. and Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), pp.1-58.

[15] Treeratpituk, P. and Giles, C.L., 2009, June. Disambiguating authors in academic publications using random forests. In Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries (pp. 39-48).

[16] Brauckhoff, D., Salamatian, K. and May, M., 2009, April. Applying PCA for traffic anomaly detection: Problems and solutions. In IEEE INFOCOM 2009 (pp. 2866-2870). IEEE.

[17] Wang, Y., Wong, J. and Miner, A., 2004, June. Anomaly intrusion detection using one class SVM. In Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004. (pp. 358-364). IEEE.

[18] Shyu, M.L., Chen, S.C., Sarinnapakorn, K. and Chang, L., 2003. A novel anomaly detection scheme based on principal component classifier. MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING.

[19] Li, K.L., Huang, H.K., Tian, S.F. and Xu, W., 2003, November. Improving one-class SVM for anomaly detection. In Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 03EX693) (Vol. 5, pp. 3077-3081). IEEE.

[20] Ma, J. and Perkins, S., 2003, July. Time-series novelty detection using one-class support vector machines. In Proceedings of the International Joint Conference on Neural Networks, 2003. (Vol. 3, pp. 1741-1745). IEEE.

[21] Hu, W., Liao, Y. and Vemuri, V.R., 2003, June. Robust anomaly detection using support vector machines. In Proceedings of the international conference on machine learning (pp. 282-289).

[22] Mahoney, M.V. and Chan, P.K., 2003, September. An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection. In International Workshop on Recent Advances in Intrusion Detection (pp. 220-237). Springer, Berlin, Heidelberg.

[23] Thottan, M. and Ji, C., 2003. Anomaly detection in IP networks. IEEE Transactions on signal processing, 51(8), pp.2191-2204.

[24] Eskin, E., Arnold, A., Prerau, M., Portnoy, L. and Stolfo, S., 2002. A geometric framework for unsupervised anomaly detection. In Applications of data mining in computer security (pp. 77-101). Springer, Boston, MA.

[25] Ye, N., Emran, S.M., Li, X. and Chen, Q., 2001, June. Statistical process control for computer intrusion detection. In Proceedings DARPA Information Survivability Conference and Exposition II. DISCEX'01 (Vol. 1, pp. 3-14). IEEE.

[26] S. M. Emran, and N. Ye, "Robustness of Canberra Metric in Computer Intrusion Detection", Proceedings of 2001 IEEE Workshop on Information Assurance and Security, US Military Academy, NY, June 2001, pp. 80-84.

[27] McHugh, J., 2000. Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Transactions on Information and System Security (TISSEC), 3(4), pp.262-294.

[28] Mukundha, C., Survey: Anomaly Detection in Cloud Based Networks and Security Measures in Cloud Date Storage Applications.

[29] G. N. Lance, and W. T. Williams. "Mixed-data classificatory programs I.) Agglomerative Systems". Australian Computer Journal, 1967, pp. 15–20.