iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** XI **Month of publication:** November 2020

DOI: https://doi.org/10.22214/ijraset.2020.32188

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XI Nov 2020- Available at www.ijraset.com

Anomaly Detection using Data Mining Techniques: A Review

Muhammad Zeeshan Younas
Department of Computer Science, Capital University of Science & Technology, Islamabad, Pakistan

Abstract: Anomaly detection is the process and technique of finding deviance or unexpected behavior in a certain dataset, this type of unforeseen behavior is also known as anomalies or outliers. Here has been talked about various kinds of anomalies and their innovative compartmentalization based on different manners. In the current world, a lot of information is stored and moved from one side to an alternative location.
When data transferred it can be hacked by attackers, whereas different methods and techniques are present to protect data from unknown attacks.
Therefore, to investigate and examine the information and to handle different types of attacks. Data mining methodologies have come out to make it minimum endangered. Different hybrid methods can be used to recognize identified and unidentified attacks more precisely.
This paper analyzed different data mining methods for anomaly detection to provide improved interpretation for researchers. The paper introduced some considerable data mining techniques used to detect anomalies.
Keywords: Anomaly Detection, Intrusion Detection System, Clustering, Data Mining, Classification.

I. INTRODUCTION

Anomaly detection is the procedure and technique of finding deviance or an uncommon fact in a certain dataset. The arrangements in which the current behavior is not normal as compare to the previous working and it can be unexpected behaviors which are previously not recognized, sometimes it may be dangerous, and sometimes it maybe not. This word anomaly is also known as an outlier. Many researchers who are working on data mining, normally there focused on some other methods like clustering and classification.

But, assessment experienced a revolution in 2000 when researchers and scholars originate recognition of unusual things that can help to resolve the actual world difficulties seen in harm detection, fraud or scam detection, detection of strange medical state, and intrusion detection. Sometimes anomalies hold appreciated information about irregular features of the systems [1]. There are three types of anomalies and it can be classified as point, contextual and collective anomaly. If a particular occurrence can be measured as anomalous in respect of its aspects, it is identified as a point anomaly. For instance, if data instance is irregular in a precise situation. The anomaly arises at a definite time or a certain region. Collective anomalies can be shaped as a group of associated data instances is irregular concerning the complete dataset, but not separate standards.

Anomaly detection is the maximum correct when it is founded on whimsicality, we can use anomaly decoction in the future as reminder support [2]. Anomaly points is the unexpected points that are all away from other points as shown in figure 1. Intrusion detection contains many outfits and methods such as statistics, machine learning, and data mining for the detection of an attack and that screen a network or system for malicious movement. Data mining techniques and methods for Intrusion Detection systems (IDS) providing maximum accuracy and admirable finding on various kinds of attacks. IDS provided support to the security of communication and advance information systems to avoid dangerous actions and security defilements [3]. To main detection procedure and to direct the exposed patterns, background information may be used to direct the exposed patterns not solitary on a brief period but at many stages of the concept.

Anomalies are the patterns in the data set that don't follow a well-organized idea of regular behavior, Anomalies can't continuously be categorized as an attack but it may be shocking or unpredicted behavior that is earlier not recognized [4]. The pattern exposed should be stimulating because moreover, they characterize mutual information or deficiency innovation. This data cannot proceed until it is converted into beneficial information, and it compulsory to analyze a large amount of data and mining useful information from it. It also examines the arrangements that diverge from predictable standards. The system should learn to categorize present behavior as reliable or abnormal with previous behavior by the only optimistic instance of the data set [5].

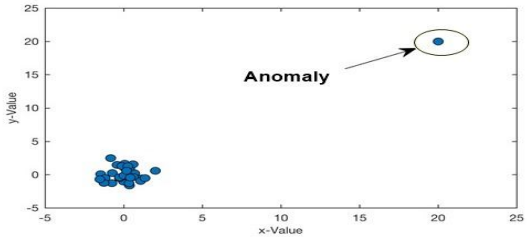


Fig. 1. Anomaly points (the points which are all away from other points).

In this work, the aim of this paper to identify and categorize a well understanding of the different kinds of data mining methods to anomaly detection continuing presently. In this work, the author attempts to associate current data mining techniques to gain improved performance outcomes, detection of irregular or unpredicted behavior will capitulate to training and classify it into innovative kinds of attacks or another specific type of intrusions. Anomaly detection in data mining is an innovative study effort that delivers the investigation of exact data by using machine learning and data mining techniques. This research work with innovative methods that are done by several researchers and the assessment will be supportive to researchers for achieving a basic vision of different methods for anomaly detection.

II. LITERATURE REVIEW

This research work presented an extensive variety of methodologies that are appropriate for anomaly detection systems in the data mining field and other social network areas. This work core aims to analyze the social networks centric anomaly detection methods that are mainly categorized as behavior-based, spectral based, and structure-based. Most of the classification additional integrates the number of methods that have conversed in the work. According to this work, the social network's current issues represent uncommon and illegal performance showing various behavior than others existing in the same arrangement [6].

Soft margin Support Vector Machine (SVM) is one of the renowned simple SVM methods using supervised learning and they relate the “one class SVM” techniques by unsupervised learning for perceiving anomalies. It means that one-class SVM is not required characterized information. They have proposed a new SVM method namely, enhanced SVM that has been used to combine these two methods for providing low false alarm competence, the same as that of a supervised SVM method [7].

Anomaly detection is the run-through of recognizing substances or actions that do not follow a projected behavior. They have proposed two novel techniques of identifying network presentation detection anomaly that is founded on split-simple arrangement namely, feedforward neural network and AdaBoost. The method tested on simulated data set for checking their understanding in respect of period and amplitude of anomaly [8]. Data mining methods are offered for probabilistic classification of oceanic traffic and anomaly detection. They have proposed a data mining technique that delivers a comparatively forthright and unsupervised method to regulate maritime rotating units and to illustrate the maritime traffic in each turning unit [9].

The core aim of this research work is to deliver an indication of numerous phases of anomaly-based intrusion detection systems. Currently, Host-based IDS (HIDS) are suitable additional significant and play a main part in the maximum of the intrusion detection systems. It has been detected that HIDS with several data mining algorithms and cluster-based methods stretch extra precise outcomes with a smaller amount of incorrect alarm rates. Their research work provided a theme of current anomaly detection methods and that how the methods can be used and applied in another application area of the domain [10].

The goal of this study is to provide a wide-ranging complete indication of the advanced approaches for anomaly fraud detection. They have used both data mining and machine learning approaches for anomaly detection. They have presented the procedure of the deviance discovery and various methods of supervised, semi-supervised, and unsupervised learning [11].

This work presented the explanation for unsupervised anomaly detection to identify unpredicted action of the operator or network devices grounded on the examination of joint dependences of the distinct portions of network movement. Subsequent model is a collaborative of fuzzy implication structures, which label the necessity of the designated limitation from other unrushed amounts standards. Anomaly detection awareness over associative analysis, the technique can identify no solitary the difficulties circumstances but also can find the maximum likely cause of the anomaly [12].

This paper discussed Recurrent Neural Networks (RNN) for identifying anomaly if flying data. RNN is also appropriate for execution on the flying deck for actual anomaly detection. RNNs design and training approach to perceive runway variation conformation and irregular pitch anomalies. Research with changing features amalgamations may be appreciated in measuring the enactment of RNN in identifying even the fine-drawn anomalies in the data set [13].

They have discussed off-line IDS and that is applied by using Multi-Layer Perceptron (MLP) artificial neural network. Knowledge Discovery in Databases (KDD) Dataset is used for the training and assessment of the ANN classifier. Upcoming work will be preceding the middle standards and directory standards attained to signify the data patterns and trained the arrangement in minimum iterations [14].

They have proposed the hybrid approach in which the clipping approach has been used to decrease the number of hyper containers and henceforth accuracy is enhanced. Different data sets are used for challenging determination. The proposed technique procedures multi-level-based, Fourier Modal Method (FMM) methods that provide improved accuracy with a minimum amount of hyper boxes using pruning strategy. This effort can be used in the future for speech arrangement and text cataloging [15].

III. TECHNIQUES AND METHODOLOGY

A. *Architecture Diagram*

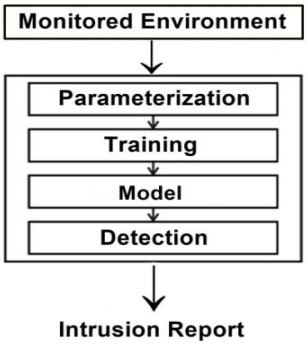


Fig. 2. The methodology of Anomaly Detection

B. *Simple Procedure of Anomaly Detection Technique*

There are Various anomaly methods sequence defined, as presented in figure 2.

1) *Training Stage:* A model is constructed at the beginning of the usual (or irregular) manners of the structure. Here various methods can be chosen dependent on the kind of anomaly detection measured. It may be automatic and may be manual.
2) *Detection Stage:* Once the model for the classification is accessible, it is associated with the framework or the pre-set detected movement. If the nonconformity originates surpasses or is fewer, when in the situation of irregularity prototypes from a preset edge then an alert will be activated.

C. *Clustering Based on Anomaly Detection Techniques*

A cluster is a group of substances that fit into a similar class. There are similar substances are gathered in a single cluster and unrelated objects are gathered in another cluster. Cluster analysis is generally used in several requests such as pattern identification, data analysis, and image processing.

1) *K-Mean:* clustering is the modest and general unsupervised ML algorithms. Normally, unsupervised algorithms mark implications from data sets using solitary input vectors deprived of mentioning to recognized, or categorized, results. The ablative reductive clustering is created on a compactness quantity intended in a given dataset fact through expending the nearest data points [16] [17].
2) *Multivariate Outlier Detection:* In some stages, multivariable interpretations do not perceive as deviation once every variable is measured self-sufficiently. Deviation recognition can be conceivable when the multi-faceted analysis is achieved, and the connections between dissimilar fluctuating are associated inside the class of data. [18].
3) *K-medoids:* The k-medoids or separating around medoids is a clustering algorithm. Both, k-means and k-medoids algorithms are partitioned and effort to reduce the space among facts categorized to be in a cluster and a fact chosen as the midpoint of that cluster [19].
4) *Subtractive Clustering Algorithm:* There core standard of the algorithm is selecting the dataset trajectories as aspirant cluster centroids. For every aspirant centroid, a compactness quantity is considered and the first centroid will be the dataset trajectory with advanced concreteness measure [16].

D. Classification Based Anomaly Detection

Classification can be used for recognizing the group of innovative examples on the foundation of a preparation established of data holding explanations or occurrences or tuples which group association is recognized. Classification is used for the train a model from a set of considered data cases. Anomaly detection methodology can be categorized into various two phases namely, the training stage and testing stage. And same as anomaly detection classifies data hooked on two main groups regular or irregular. There are the following ML techniques and methods in anomaly detection

1) *Neural Networks:* It is a significant module of lenient computing methods that have been established with the new inspiration of the info dispensation style of the human brain and it is stimulated by the biological neural networks system. NN is understood as corresponding computation simulations and contains parallel and comprehensive execution of nonlinear static or dynamic systems. Multilayer Perceptions (MLP) of neural networks is effective in the variability of applications and generating more correct outcomes than other current computational learning techniques [20].
2) *Support Vector Machine:* SVM is a perform with solid regularization method, the rationalization approach increases the revelatory precision while certainly holding away from over-fitting of the training information. The major work of the SVM is to find the hyperplane that divided them into two classes. It is very effective in high-dimensional spaces and its work is very impressive in a clear margin of separation. SVM does not perform well when we have a large amount of dataset. Its performance is very is low when data is noisy [21].
3) *Classification Tree:* A classification tree in machine learning methodology is also known as a decision tree or prediction model. In which some tree pattern type graphs are similar to flow chart structure. The goal is to construct a model that can predict the worth of a marked variable by learning modest decision rules. There is the most common and fundamental algorithms used for classification trees are C4.5 and ID3. It is very simplest to understand and recognize. And it can be sometimes unstable [22].
4) *Fuzzy Logic:* Fuzzy logic derived from a fuzzy set theory which deals with the cognition that is estimated rather than exactly assumed from classical establishes logic. In this method, the data set is categorized based on different statistical metrics. Data set is applied with fuzzy logic rules and it generates some results as normal or malicious. The fuzzy logic system can take distorted and noisy inputs of information. But there is no systematic approach to fuzzy system design [23].
5) *Rule-Based:* In a Rule-based algorithm every classification technique uses a procedure to produce rules from the model data. These rules are then applied to fresh data. It intent on an exact class at a time and exploiting the probability of the requirements classification. Rule enchanting into explanation network movement, many of intrusion or Trojan horse is based on the difference of the unique perfect. As we know, Red Code1 and Red Code2 they both are caterpillar which is attacks that spread themselves over networks deprived of any user interference or interaction [24].
6) *Naïve Bayes Network:* Naive Bayesian classifiers in a method that they consent dependencies among attribute values to be defined. It outpaces the good performance because of the clear demonstration of the fundamental structure and the existence of human proficiency knowledge thus accumulative the learning rat. It uses the concept of provisional probability. It needs a small amount of training data set and is not suitable for large datasets [6].

E. Hybrid Approaches

By using any particular approach, we cannot yield proper and appropriate outcomes. A single algorithm is not sufficient to detect new attacks that are newly introduced gradually. Combination and merging of various algorithms together have been used in the past few years for getting proper and accurate results.

1) *Combining Supervised and Unsupervised Techniques:* There are many supervised and unsupervised learning algorithms whose mixture can be prepared. In the current previous duration, various hybrid techniques are approached. Through this, the performance of a supervised algorithm is extremely improved as the accuracy of anomaly detection percentage can be highly upgraded by implemented unsupervised algorithms. The hybrid methodology is used for detecting anomalies in the network which is a mixture of together entropy and SVM approaches. As a problem by modest entropy-based technique is stable threshold variety for entropy is static for classifying anomaly detection. This process is not an exact dynamic to select whether there is an attack, because entropy standards can diverge from the fixed sort in standard conditions. By using the SVM model individually we cannot get good results as network features are used for learning without dispensation. Experimental outcome validates that this hybrid process works well with maximum accuracy for recognition of attack traffic and fewer false alarms [25]. Consequently, hybrid methods yield enhanced outcomes as merging various techniques.

2) *Cascading Supervised Techniques:* Some different classification algorithms are engaged together to get maximum accuracy and a grouping of decision tree and naïve Bayes algorithms was suggested. This hybrid technique was tried in Knowledge Data Discovery (KDD) of the data set and the accuracy accomplished was 99 %. It focused on the growth of the working of the Naïve Bayesian (NB) classifier and the ID3 algorithm [26]. There are numerous kinds of mixtures that are thinkable therefore several methods can be suggested and the finest resulting methods can be applied essentially.

IV. COMPARITIVE STUDY

Table 1. Compendium of hybrid methodologies for anomaly detection

Methods Used	Authors Name	Methodology Implementations	Advantages & Disadvantages.
Naïve Bayes Classification and k-Medoids Clustering	Hae-Sang Park and Chi-Hyuck Jun [19].	Same data occurrences are assembled by using the K-Medoids clustering method. Resulting clusters are categorized using Naïve Bayes classifiers.	Inflexible to predict when naïve bayes classifier in dissimilar situations. Maximum in recognition amount and decrease in for wrong alarm amount.
K-means and Subtractive clustering algorithm	István Kiss and Béla Genge [16].	Numerous clustering methods are discovered to indicate the maximum appropriate for clustering the data points indexed features, therefore categorizing the conditions and possible cyber-attacks to the physical system.	Overtakes the isolated K-Means and the ID-3. This method is incomplete to exact Data set. Different initial centroids can result in various clusters.
Neural Network (MLP) and Naive Bayes	Ravneet Kaur and Sarbjeet Singh [6].	Anomalous actions in social networks signify strange and prohibited activities. IDS applied by MLP artificial NN. The minimum error rate is produced by Bayesian classifiers and it is more accurate. .	The survey of a quantity of additional chart benchmark that can be utilized to identify the novel types of anomalies current in various community networks. Does not work accurately if their features are correlated, not suitable for a large dataset.
Decision Tree (DT)	Wu S. Y. and Yen E [22].	The interior bulges are assessment stuff, individually branch signifies test outcomes, and the maximum important or mutual algorithm utilized for classification tree is ID-3 and C4-5.	Simple to understand and to interpret and requires small data preparation. Decision trees can be unstable and disposed of overfitting.
One-Class and Two-Class Support Vector Machines (SVM)	Murthy D, Gross A, and Takata A [21].	One-class SVM is utilized for identifying irregular marks. Furthermore, a two-class indicator is re-educated once convinced novel information proceedings are comprised in the present data set.	It doesn't want a previous disenchantment and is convertible through gain knowledge from experiential loss proceedings. The accurateness of non-success recognition can't influence 99.99%. It is effective in a situation where the number of dimensions is higher than the number of samples.

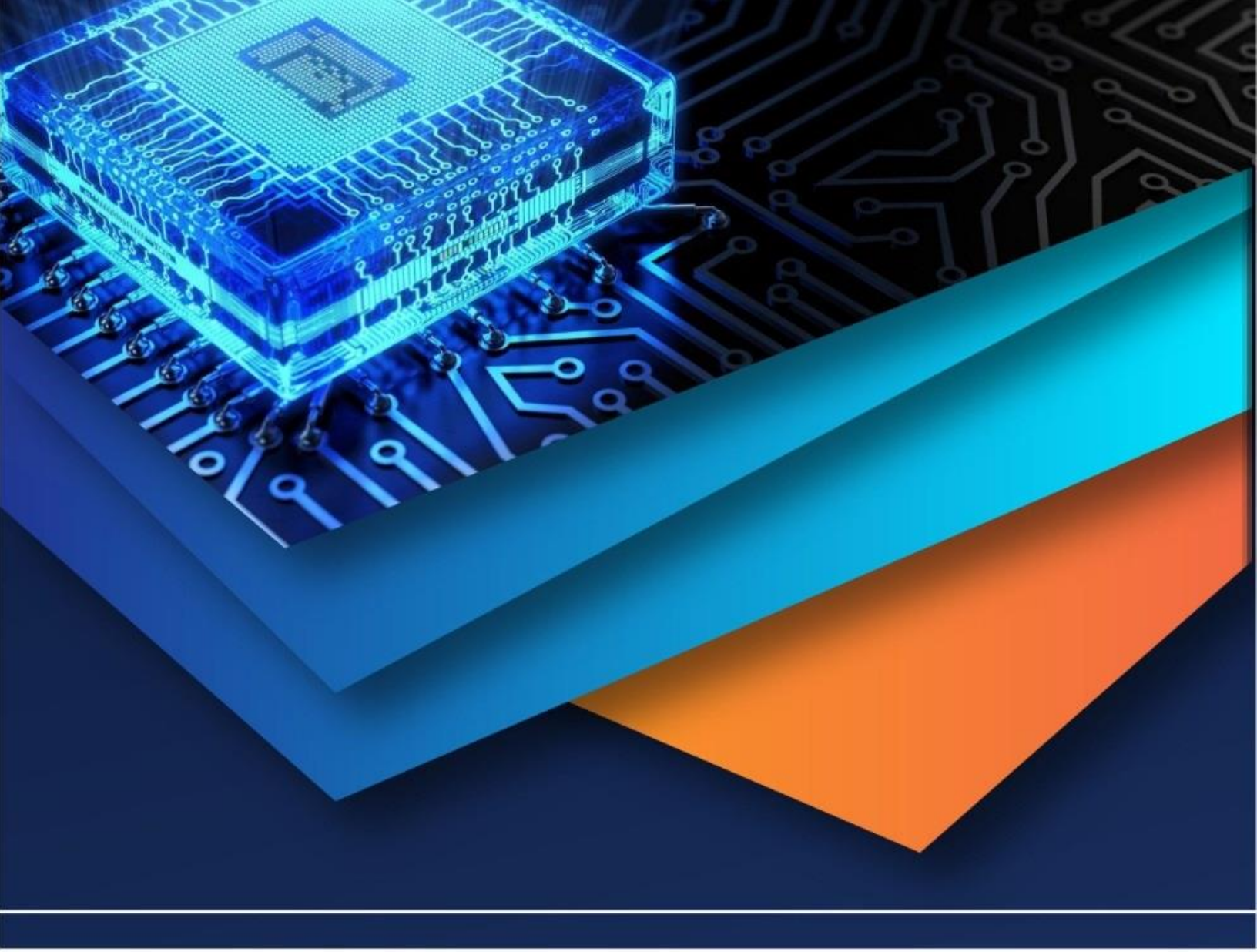
V. DISCUSSION AND CONCLUSION

In this work, the aim of this paper to identify and categorize a well understanding between the different kinds of data mining methods to anomaly detection that is continuing presently and that has been proposed for the past some years. This assessment will be supportive to researchers and scholars for achieving a basic understanding of several methodologies for anomaly detection. Even though the considerable effort had been performed by using, hybrid approaches and independent algorithms those are in existence massively used as they deliver well outcomes and results in the disadvantage of one method over the other. By using any particular approach, we cannot yield proper and appropriate outcomes. A single algorithm is not sufficient to detect new attacks that are newly introduced gradually. Combination and merging of various algorithms together have been used in past few years for getting proper and accurate results. Diurnal the fresh unidentified occurrences are observed and there is an essential requirement of those methodologies that can identify the unfamiliar behavior in the dataset safe transported or improved. This work with innovative methods that are done by several researchers and the review would be supportive to scholars and researchers for achieving a core vision of different methods for anomaly detection. This work combination or a mixture of previously prevailing algorithms is declared that have been suggested. Attentive researchers can associate the changed form of by now current algorithms. For instance, there are different new methodologies in the alteration of the decision tree for example C4-5 and ID-3. Containing enhanced and numerous core-based methodologies GA, SVM. This may produce more precise outcomes. In the coming time, we can utilize the method and achieve a qualified analysis with clustering methods for the time sequence information using ANN.

REFERENCES

[1] Anitha Ramchandran and Arun Kumar Sangaiah, “Unsupervised Anomaly Detection For High Dimensional Data—An Exploratory Analysis”, Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications, 2018 (pp.233-251). DOI: 10.1016/B978-0-12-813314-9.00011-6.
[2] Jakkula and D.J. Cook, “Anomaly Detection Using Temporal Data Mining in a Smart Home Environment”, 2008 Methods of Information in Medicine 47(1) pp 70-5, DOI: 10.3414/ME9103.
[3] Roma Sahani, Shatabdinalini, “Classification of Intrusion Detection Using Data Mining Techniques”, Classification of Intrusion Detection Using Data Mining Techniques, 2018. doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-981-10-7871-2_72
[4] Jagruti D. Parmar and Jalpa T. Patel, “Anomaly Detection in Data Mining: A Review”, International Journal of Advanced Research in Computer Science and Software Engineering, 2017. Vol 7(4).
[5] Terran Lane and Carla E. Brodley, “An Application of Machine Learning to Anomaly Detection”, In Proceedings of the 20th National Information Systems Security Conference, 1997.
[6] Ravneet Kaur and Sarbjeet Singh, “A survey of data mining and social network analysis based anomaly detection techniques”, Egyptian Informatics Journal, 2016. Vol 17(2) pp199-216, DOI: doi.org/10.1016/j.eij.2015.11.004.
[7] Taeshik Shon and Jongsub Moon, “A hybrid machine learning approach to network anomaly detection”, 2007 Information Sciences 177(18) pp 3799-3821, DOI: 10.1016/j.ins.2007.03.025.
[8] James Zhang, Robert Gardner and Ilija Vukotic, “Anomaly detection in wide area network meshes using two machine learning algorithms”, Future Generation Computer Systems 2018. DOI: https://doi.org/10.1016/j.future.2018.07.023.
[9] H. Rong, A.P. Teixeira and C. Guedes Soares, “Data mining approach to shipping route characterization and anomaly detection based on AIS data”, Ocean Engineering 2020, Vol 198, 106936, DOI: 10.1016/j.oceaneng.2020.106936.
[10] Shijoe Jose, D.Malathi, Bharath Reddy and Dorathi Jayaseeli, “A Survey on Anomaly Based Host Intrusion Detection System”, National Conference on Mathematical Techniques and its Applications (NCMTA 18) 2018, DOI: 10.1088/1742-6596/1000/1/012049.
[11] Leman Akoglu and Christos Faloutsos, “Anomaly, Event, and Fraud Detection in Large Network Datasets”, WSDM '13 2013, DOI: 10.1145/2433396.2433496.
[12] Tetiana Gladkykh, Taras Hnot and Volodymyr Solskyy, “Fuzzy Logic Inference for Unsupervis Anomaly Detection”, IEEE First International Conference on Data Stream Mining & Processing 2016. pp 42-47.
[13] Anvardh Nanduri and Lance Sherry, “Anomaly Detection In Aircraft Data Using Recurrent Neural Networks (RNN)”, IEEE Integrated Communications Navigation and Surveillance (ICNS) Conference, 5C2-8(2016):19-21.
[14] Swain Sunita, Badajena J Chandrakanta and Rout Chinmayee, “A Hybrid Approach of Intrusion Detection using ANN and FCM”, European Journal of Advances in Engineering and Technology, 3(2), (2016): 6-14.
[15] Bhavana Jain and Vaishali Kolhe, “Hybrid Approach for Classification using Multilevel Fuzzy Min-Max Neural Network”, International Journal of Innovative Research in Computer and Communication Engineering, Volume 4, Issue 5 (2016): 8636-8640.
[16] István Kiss and Béla Genge, “Data Clustering-based Anomaly Detection in Industrial Control Systems”, Conference Intelligent Computer Communication and Processing 2014, DOI :10.1109/ICCP.2014.6937009.
[17] Amuthan Prabakar Muniyandia, R. Rajeswarib and R. Rajaramc, “Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm, International Conference on Communication Technology and System Design 2011. DOI: 10.1016/j.proeng.2012.01.849.
[18] Irad Ben-Gal, “OUTLIER DETECTION”, Data Mining and Knowledge Discovery Handbook 2005, pp 131-146, DOIhttps://doi.org/10.1007/0-387-25465-X_7.
[19] Hae-Sang Park and Chi-Hyuck Jun, “A simple and fast algorithm for K-medoids clustering”, Expert Systems with Applications 36 (2009) 3336–3341, DOI: doi:10.1016/j.eswa.2008.01.039.
[20] Ferdi Sönmez, Metin Zontul, Oğuz Kaynar and Hayati Tutar, “Anomaly Detection Using Data Mining Methods in IT Systems: A Decision Support Application”, Sakarya University Journal Of Science 2018, DOI: 10.16984/saufenbilder.365931.

[21] Murthy D, Gross A, Takata A and Bond S, "Evaluation and Development of Data Mining Tools for Social Network Analysis", In Mining Social Networks and Security Informatics, Springer 2013. pp. 183-202.
[22] Wu S. Y., Yen E, "Data mining-based intrusion detectors", Expert Systems with Applications; 36(3); 2009; p. 5605-5612.
[23] Kaur N, "Survey paper on Data Mining techniques of Intrusion Detection", International Journal of Science, Engineering and Technology Research 2013, pp 799-804.
[24] Lei Li, De-Zhang Yang and Fang-Cheng Shen, "A Novel Rule-based Intrusion Detection System Using Data Mining", Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference, Vol 6, DOI: 10.1109/ICCSIT.2010.5563714.
[25] Basant Agarwal and Namita Mittal, "Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques", 2nd International Conference on Communication, Computing & Security (ICCCS-2012), DOI:10.1016/j.protcy.2012.10.121
[26] 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 2010), pp 12-25, DOI:10.5121/ijnsa.2010.2202.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY