A Review on Outlier Detection for Data Mining using Machine Learning

Himanee Mishra[1], Chetan Gupta[2], Dr. Ritu Shrivastava[3]
M.TechResearch Scholar,Dept. ofCSE[1], Assistant Professor, Dept. of CSE[2], HOD Dept. of CSE[3]
mishra.himanee@gmail.com[1], chetangupta.gupta1@gmail.com[2,] ritushrivastava08@gmail.com[3]
SIRTS, Bhopal, India[1], SGI, Bhopal, India[2], SGI, Bhopal, India[3]

Abstract: -OD is a Data Mining Application. Anomaly contains boisterous information which is explored in different areas. The different strategies are now being explored that is more conventional. We reviewed on different procedures and uses of OD that gives an original methodology that is more helpful for the novices. The proposed approach assists with cleaning information at college level significantly quicker with incredible precision with the assistance of AI calculation. This study incorporates the current exception strategies and applications where the loud information exists. Our paper characterizes basic audit on different methods utilized in various utilizations of exception recognition that are to be investigated further and they gives a specific sort of information based information for example more helpful in research exercises. So where the Anomalies are available it will be distinguished through OD methods and checked in like manner.

Keywords: -Outlier Detection (OD), Data Mining, Machine Learning (ML)

I. INTRODUCTION

Data mining is an eminent process to determine information, like patterns, associations, variations and substantial structure from large amount of data kept in databases. On the account of comprehensive accessibility of enormous amount of data in electronic form and required for turning such data into useful information as well as for knowledge broadcast applications, that includes marketing, business management and support on decision making. Recently, the data mining attracts an abundant deal of attention in information industry. It is an essential process which exhibits intelligent scheme to extract data patterns. Data base technology is familiar for storing and exploring large volumes of data [1]. Data mining also use to discover their mutual relationships. Data mining algorithms typically necessitate various passes over the large amount of data and most of them are computationally demanding. Powerful data mining tools are necessary for the several trends such as an increasingly service based company, the growth of mass communication, the competing importance of the information. It is a process of interactive and iterative method consisting of many steps, selection of data, its pre-processing, transformation, adequate mining and interpretation of secure results [2, 3]. The scope of data mining brings together ideas and techniques from the variety of field's economics, artificial intelligence, data bases and statistics. The task of data mining also improves its marketing, sales, operations that supports the customer over better understanding of its customers. The process of knowledge discovery is comprised of steps that involve iterative order such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, knowledge presentation etc. Data mining deals with finding the hidden patterns and non-trivial conditions from different types of data. Outlier detection and future prediction are the important problem in data mining technology. Outlier detection is the process of finding outliers depending up on the behavior and distribution of data. Outliersareanomalies that differ from other data points in the dataset. The main intension of outlier detection is to identify the atypical data with unpredictable features that are produced by various mechanisms. The outlier in a dataset also consider as form of clusters CN1 and CN2 denotes the normal behavior and the data O2 and O3 has different behavior than the cluster CN1 and CN2.Although the cluster O1 has some data points, it is considered to be an outlier. Nowadays, the detection of bank transaction has gained popularity in recent years. There are several existing methods for outlier detection namely rule based approach, clustering based approach, classification-based approach and model-based approach. However, the existing techniques detect the outliers in the financial transactions, they are not capable of handling a larger volume of data [4, 5].

II. OUTLIER DETECTION IN DATA MINING

In the dataset, the outlier is an anomaly that differs from the different data points. In data mining it is also termed as abnormalities, deviants, abnormalities and anomalies. In data mining, the major problems occur are the outlier detection and future prediction techniques. It is defined as the process of finding outliers that depending up on the behavior and distribution of data. The discoveries of abnormal features with inconsistent characteristics are one of the intension of outlier detection. The regression modelling, removes the outlier method and it considers separately to improve the accuracy [6]. There are various categories of outliers namely, Point outliers, Context outliers, Collaborative outlier, vector outlier, sequence outlier and graph outlier. The primary step of data mining applications is the outlier detection There are numerous methods associates with outlier detection, such as differentiating amongst the univariatevs multivariate techniques and parametric vs non-parametric measures. If outliers carry some information, it may consider as error or noise. Outlier detection methods suggests for various applications such as detection of fraudulent in credit card, voting irregularity examination, medical trials, data cleaning, network interruption, severe weather forecast, environmental information and other data mining tasks. There are various difficulties in analyzing an outlier in its present form [7]. Encompassing every possible normal behavior in the region. The normal and outlier have an imprecise boundary. It is difficult to determine and remove, due to noise in the data. Due to the contrary notion of outliers, it is difficult to apply the technique in one domain to another domain [8].

Outlier Detection Methodologies

1. Statistical-Based OD
2. Deviation-Based OD
3. Distance-Based OD

Challenges of OD

1. Displaying common substances and exceptions suitably
2. Extreme to appraise all probable commonplace exercises in an application The limit among the average and exception substances is normally an defined situation
3. Application-explicit exception discovery Selection of distance measurein the midst of substances and the model of association between the substances are frequently application-dependent.
4. Tackling noise in outlier detection
5. Noise may falsify the typical substances and distort the difference amongst the typical substances and outliers. It helps to hide the outliers and minimizes the efficacy of outlier detection
6. Understandability Understanding the outliers: Validation of the detection Identify the grade of an outlier: the impossibility of the item being made by a standard strategy.

III. TYPES OF OUTLIERS

A vital part of an exception recognition method is the idea of the ideal anomaly. Exception Classification is done based on their event; for the most part there are three sorts of anomalies which are counted as follows [9, 10]:

1. Point Outliers
2. Context oriented Outliers
3. Aggregate Outliers

Point Outlier: when an information occurrence is not the same as set of information then, at that point, occasion is named as point anomaly. It is the most straightforward type of exception and utilized in different explores. For instance Visa misrepresentation Detection, the anomaly can be identified concerning sum spent assuming consumption is higher contrasted with ordinary exchanges then it is an exception [11].

Context oriented Outlier: when an information occurrence is abnormal regarding some unique situation (condition), then, at that point, case said to be Contextual Outlier. Context oriented exceptions for the most part investigated on time series information. For instance, in setting old enough a six feet grown-up might be a typical individual while six feet kid is an exception [12].

Aggregate Outlier: When an assortment of related information is irregular from rest of the whole informational index, then, at that point, it is an aggregate Outlier. They can happen just in informational indexes where information occurrences are connected. Aggregate anomaly has been investigated on graphical information, consecutive information and spatial information. For instance Human Electrocardiogram yield showed in Figure 1. The featured district is an anomaly since same worth exists for the unusually lengthy timespan. Aggregate anomalies can be applied for chart information, grouping information and spatial information [13].

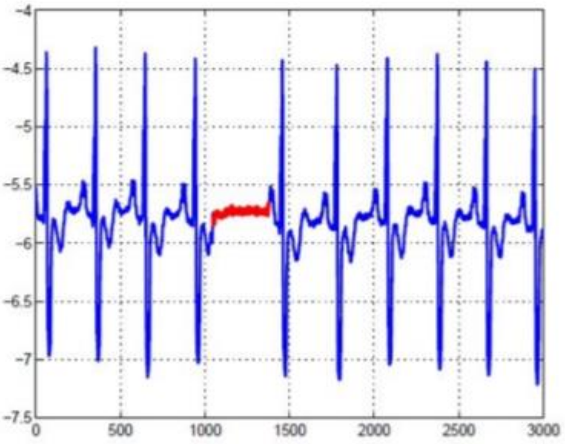

Fig. 1: Collective Outlier

IV. RELATED WORK

Outlier detection is defined as the process of finding outliers based on the behavior and distribution of data. The main intention of outlier detection is to recognize the abnormal data with unpredictable features that are produced by various mechanisms.

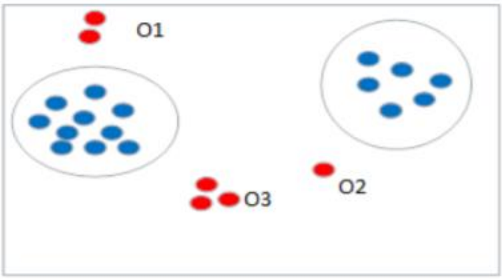

Fig. 2: outliers in dataset

Generally, the outlier is an opinion of the data that can be removed or considered separately in regression modeling to improve the accuracy. Moreover, it is a pattern of data that do not approve to a distinct notion of normal behavior.

In ref. [1], suggested a anomaly detection was an essential task in data mining that have an intention to found components to show the different behavior compared to the other forms and such were known as outliers. One amongst the broadly utilized measures for determining whether a component was an exception depends on the sum of neighboring components within a distance and a limit. That types of outliers were removed based on the exceptions. It states both an extendible structure for anomaly identification calculations and exact exception location calculations were mentioned, where the anomaly recognition was persistently completed over an information stream.These algorithms were realized in MOA framework, for extend its process with outlier detection proficiencies. One more significant problem for additional exploration was the capability for visualizing the data streams in any metric space. The multidimensional data sets were reinforced (specifically, 2D data sets) for visualization limitations. This challenge was incorporated into MOA operative perception of common metric spaces.

In ref. [2], recommended an Improved Genetic KMeans (IGK) algorithm to proficiently discover the outliers. The main concern of this clustering-based outlier detection algorithm was used to identify the outliers and data clustering concurrently. In this, it regularly observed the noise, which should be detached in order to provide more reliable clustering method. During the process of clustering and outlier discovery they estimate the centroids of the generative distribution process. The IGK was an efficient clustering technique that handles the large amount data with the help of Genetic Algorithm (GA). The findings of this technique were as used to: avoid the disserving big clusters. In some degree, it overwhelmed the deflection of data and it reduced the sensitivity to isolated point. Here, the outlier detection could be accomplished only on numeric dataset. When the clustering method was utilized in outlier detection, then they focused mainly on those elements as outliers which was deceitful outside the clusters.

In ref. [3], described the basic process of outlier detection in data mining tasks were well examined because of its various applications.In that most applications occur in high-dimensional spaces. A blockage of prevailing methodologies was certain or, on the other hand unambiguous evaluations on ideas of separation or closest neighbor were decayed in highdimensional data. The utilization of angle-based outlier aspect was considered in mining high-dimensional outliers. That technique run in cubic time with a quadratic time heuristic, they suggest a new irregular projection-based strategy that can assess the edge-based outlier aspect for the data which were focused on time linear in the data size. Additionally, their method was reasonable to perform in corresponding condition to accomplish an equivalent speedup. They offered a hypothetical study of the quality estimation to ensure the fixed quality of this proposed assessment system. The observational tests on manufactured and real-world datasets exhibit the scalability, efficacy and competence to identify the outliers in large high-dimensional data sets.

In ref. [4], introduces an data mining (DM) based way to deal with creating outfit models for anticipating following day vitality utilization and pinnacle control request, with the point of enhancing the forecast precision. In addition, an outlier detection method was also offered to detect the abnormal building operative patterns. It was more implemented for analyzing the huge energy consumption data of the highest building. Three different stages were involved in this approach, they were: Right off the bat, exception location, which consolidates highlight extraction, grouping examination, and the summed up extraordinary studentized veer off (GESD), was performed to evacuate the anomalous day by day vitality utilization profiles. Also, the recursive element end (RFE), an installed variable determination strategy, was connected to choose the ideal contributions to the base expectation models grown independently utilizing eight famous prescient calculations. The outcomes additionally demonstrate that the exception identification strategy has powerful in recognizing the rare day by day energy utilization profiles. The RFE procedure can fundamentally diminish the calculation stack while upgrading the model execution. The gathering models were important for creating methodologies of fault identification and determination in advance operation. The multiple linear regression (MLR) and ARIMA models, do not execute splendidly, since the building associated processes were usually nonlinear and difficult.

In ref. [5], discussed about the exposure of distance-based outliers from huge dimensions of data stream was dangerous for current applications extending from credit card fraud detection to moving object monitoring. It considered a framework to tackle three different classes of distance-based outliers in the streaming atmospheres. They were: Minimal Probing standard used a lightweight analytical operation together least yet adequate proof for outlier detection. Lifespan-aware prioritization rule use the fleeting connections among stream information focuses to organize the handling request among them amid the testing process. Guided by these two standards, we plan an exception identification methodology which was ended up being ideal in the costs of CPU expected to decide the exception status of any information point amid its whole life. Scalability was needed to improve in modern distributed multi-core clusters of machines for outlier detection.

Problem Formulation:-

The following problems are identified in the paper-

1. The number of false positives in the algorithm is quite high in some specific cases, which can be further reduced.
2. Time complexity is high if dimension of data increases.
3. The result is undesirable when processing high-dim data.
4. Algorithmfor clustering and type3 outliers need to be researched.

V. OUTLINES DETECTION USING MACHINE LEARNING

Uproarious information is available in the heap of substance that will be identified through the anomaly strategies. The information can be spatial or can be a transient method spatial connected with the geological conditions and worldly connected with the time perspectives [14, 15]. The principle point of exception identification is to deal with the loud information that is introduced in the heap of text. Different methods for recognizing abnormalities in Text are specified in below:

Learning

The main property of an ML is its capability to learn. Learning or preparing is a procedure by methods for which a neural system adjusts to a boost by making legitimate parameter modifications, bringing about the generation of wanted reaction. Learning in an ML is chiefly ordered into two classes as [16][19].

- Supervised learning
- Unsupervised learning

Supervised Learning

Regulated learning is two stage forms, in the initial step: a model is fabricated depicting a foreordained arrangement of information classes or ideas. The model developed by investigating database tuples portrayed by traits. Each tuple is expected to have a place with a predefined class, as dictated by one of the qualities, called to have a place with a reclassified class, as controlled by one of the traits called the class name characteristic. The information tuple are dissected to fabricate the model all things considered from the preparation dataset [17].

Unsupervised learning

It is the kind of learning in which the class mark of each preparation test isn't knows, and the number or set of classes to be scholarly may not be known ahead of time. The prerequisite for having a named reaction variable in preparing information from the administered learning system may not be fulfilled in a few circumstances.

Data mining field is a highly efficient techniques like association rule learning. Data mining performs the interesting machine-learning algorithms like inductive-rule learning with the construction of decision trees to development of large databases process. Data mining techniques are employed in large interesting organizations and data investigations. Many data mining approaches use classification related methods for identification of useful information from continuous data streams.

Nearest Neighbors Algorithm

The Nearest Neighbor (NN) rule differentiates the classification of unknown data point because of closest neighbor whose class is known. The nearest neighbor is calculated based on estimation of k that represents how many nearest neighbors are taken to characterize the data point class. It utilizes more than one closest neighbor to find out the class where the given data point belong termed as KNN. The data samples are required in memory at run time called as memory-based technique. The training points are allocated weights based on their distances from the sample data point. However, the computational complexity and memory requirements remained key issue. For addressing the memory utilization problem, size of data gets minimized. The repeated patterns without additional data are removed from the training data set [18].

Naive Bayes Classifier

Naive Bayes Classifier technique is functioned based on Bayesian theorem. The designed technique is used when dimensionality of input is high. Bayesian Classifier is used for computing the possible output depending on the input. It is feasible to add new raw data at runtime. A Naive Bayes classifier represents presence (or absence) of a feature (attribute) of class that is unrelated to presence (or absence) of any other feature when class variable is known. Naïve Bayesian Classification Algorithm was introduced by Shinde S.B and AmritPriyadarshi (2015) that denotes statistical method and supervised learning method for classification. Naive Bayesian Algorithm is used to predict the heart disease. Raw hospital dataset is employed. After that, the data gets preprocessed and transformed. Finally by using the designed data mining algorithm, heart disease was predicted and accuracy was computed.

Support Vector Machine

SVM are used in many applications like medical, military for classification purpose. SVM are employed for classification, regression or ranking function. SVM depends on statistical learning theory and structural risk minimization principal. SVM determines the location of decision boundaries called hyper plane for optimal separation of classes as described in figure 3. Margin maximization through creating largest distance between separating hyper plane and instances on either side are employed to minimize upper bound on expected generalization error. Classification accuracy of SVM not depends on dimension of classified entities. The data analysis in SVM is based on convex quadratic programming. It is expensive as quadratic programming methods need large matrix operations and time consuming numerical computations [20][21].

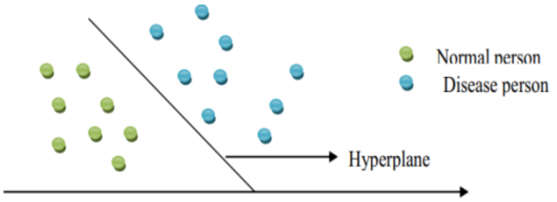

Fig. 3: Support Vector Classification

VI. CONCLUSION

We presume that basic examination on uses of exception recognition will help in additional exploration draws near. Exception data is exceptionally helpful when information is contrasted and the first information. The above basic audit will help in the further examination. Exception recognition approaches gives a basic and substantial result for the given information. Our exploration work remembers the basic examination for the different application areas and methods of the exception identification. It has been an extraordinary work for the individuals who need to begin the exploration on anomaly discovery and its space. The whole work comprises various stages and loads of hypothetical ideas in regards to the Anomalies.

We aim to propose new solutions that overcome aforementioned challenges in streaming context and adopt the sliding window technique, but efficiently store in memory a statistical summary of obsolete data, which contributes to the prediction of future data.

REFERENCES

[1] Kangqing Yu, Wei Shi and Nicola Santoro, “Designing a Streaming Algorithm for Outlier Detection in Data Mining—AnIncrementa Approach”, Sensor, MDPI 2020.

[2] Yu, K.; Shi, W.; Santoro, N.; Ma, X. Real-time Outlier Detection over Streaming Data. In Proceedings of the IEEE Smart World Congress (SWC 2019), Leicester, UK, 19–23 August 2019.

[3] Wang, Q.; Luo, Z.; Huang, J.; Feng, Y.; Liu, Z. A novel ensemble method for imbalanced data learning: Bagging of extrapolation-SMOTE SVM. Comput. Intell. Neurosci. 2017, 2017, 1827016.

[4] Hanifah, F.; Wijayanto, H.; Kurnia, A. SMOTEBagging Algorithm for Imbalanced Dataset in Logistic Regression Analysis (Case: Credit of Bank X). Appl. Math. Sci. 2015, 9, 6857–6865.

[5] Tantithamthavorn, C.; Hassan, A.; Matsumoto, K. The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. arXiv 2018, arXiv:1801.10269.

[6] Quinn, J.A.; Sugiyama, M. A least-squares approach to anomaly detection in static and sequential data. Pattern Recognit. Lett. 2014, 40, 36–40.

[7] Tran, L.; Fan, L.; Shahabi, C. Distance-based Outlier Detection in Data Streams. Proc. VLDB Endow. 2016, 9, 1089–1100.

[8] Sadik, S.; Gruenwald, L. DBOD-DS: Distance based outlier detection for data streams. In Proceedings of the 21st International Conference on Database and Expert Systems Applications (DEXA), Bilbao, Spain, 30 August–3 September 2010; pp. 122–136.

[9] Sadik, S.; Gruenwald, L. Online Outlier Detection for Data Streams. In Proceedings of the 15th Symposium on International Database Engineering & Applications Symposium (IDEAS), Lisboa, Portugal, 21–23 September 2011; pp. 88–96.

[10] Sadik, S.; Gruenwald, L.; Leal, E. In pursuit of outliers in multi-dimensional data streams. In Proceedings of the 4th IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 512–521.

[11] Forrest, S., Esponda, F., and Helman, P. 2004. Aformal framework for positive and negative detectionschemes. In IEEE Transactions on Systems, Man and Cybernetics, Part B. IEEE, 357 - 373.

[12] Gonzalez, F. A. and Dasgupta, D. 2003. Outlierdetection using real-valued negative selection.GeneticProgramming and Evolvable Machines 4, 4, 383- 403.

[13] Lee, W., Stolfo, S. J., and Mok, K. W. 2000. Adaptiveintrusion detection: A data mining approach. ArtificialIntelligence Review 14, 6, 533 – 567.

[14] Gwadera, R., Atallah, M. J., and Szpankowski, W. 2005b.Reliable detection of episodes in event sequences. Knowledge and Information Systems 7, 4, 415 - 437.

[15] Chow, C. and Yeung, D.-Y. 2002. Parzenwindown network intrusion detectors. In Proceedings of the 16thInternational Conference on Pattern

Recognition.Vol. 4.IEEE Computer Society, Washington, DC, USA, 40385.

[16] Lin, J., Keogh, E., Fu, A., and Herle, H. V. 2005. Approximations to magic: Finding unusual medical time series. In Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society, Washington, DC, USA, 329 - 334.

[17] Desforges, M., Jacob, P., and Cooper, J. 1998. Applications of probability density estimation to the detection of abnormal conditions in engineering.In Proceedings of Institute of Mechanical Engineers. Vol. 212. 687 - 703.

[18] Yairi, T., Kato, Y., and Hori, K. 2001. Fault detection by mining association rules from house-keeping data. In In Proceedings of International Symposium on Artificial Intelligence, Robotics and Automation in Space.

[19] Mr. Rahul Kumar, Chetan Gupta, Dr. RituShrivastava, "Survey on Handwritten Text Recognition Using Different Deep Learning Techniques", "International Journal of Engineering Technology and Applied Science”, (ISSN: 2395 3853), Vol. 7 Issue 11 November 2021.

[20] Ms. Tanushri Jain1, ChetanGupta,"A Review on Intrusion Detection System using Deep Learning", "International Journal of Creative Research Thoughts", IJCRT Volume 8, Issue 7 July 2020.

[21] Soni S, Dubey S, Tiwari R, Dixit M. Feature Based Sentiment Analysis of Product Reviews Using Deep Learning Methods. International Journal of Advanced Technology & Engineering Research (IJATER). 2018.