OriginalResearchArticle

J. Computing & Int. Systems (2022)270–273

SACRED HEART RESEARCH PUBLICATIONS

Journal of Computing and Intelligent Systems

Journal homepage:www.shcpub.edu.in

ISSN: 2456-9496

Hit of Fit: A Moral Analysis of Detecting Outliers using Clustering-Based Approach

Rajalakshmi[#1], P. Madubala[#2]

Received on 25 APR 2022, Accepted on 15 JUN 2022

Abstract —Researchers are becoming increasingly interested in odd behaviour as machine learning and data analytics develop. The goal of this work is to improve the estimation of numerous factors that hide outliers and produce the best clustering results. In the real-time environment, outlier detection is an important research area. To deal with the overwhelming amount of uncertainty, the following factors make the approach difficult: a) defining the boundary between normal and abnormal behaviour b) cluster size distributed over space; c) stabilising the centroid point; d) model validation; e) difficult to distinguish noise; f) modifying unnormalized data into robust data; g) estimating fuzzy function point metric that gives more insight to imbalanced data in outliers using fuzziness. The primary goal of the first stage is to reduce noise. The data is clustered using fuzzy clustering in the second stage. To stabilise the centroid, the third stage uses a variety of algorithms, including numerous rounds of k-means, fuzzy clustering, and an improved method. After that, determine the fuzzy-fp metric, fuzzy variable index, and utilisation factor to complete the outcome analysis. Using the silhouette, Pseudo-F-Statistic, and Constellation methods, the cluster goodness is verified in the last stage. The benefits and drawbacks of this strategy are effectively listed. In a huge dataset, the outcome reveals a higher performance that identifies the factor by a high performance score and adds insight to provide a balanced view of the data.

Keywords– Clustering and Fuzzy.

I. INTRODUCTION

In a variety of sectors, data mining is becoming increasingly popular. Researchers are tremendously interested in uncovering unexpected behaviour across large datasets. In data mining, outlier detection is intensively investigated and developed for specific application domains, while others are generic in nature. It's one of the most significant and hotly debated topics in science, and it's facing a slew of new difficulties. Changes in system behaviour, mechanical faults, human error, natural variations, and instrumental error all includes in it. Due to technological advancements, it aids in the comprehension of numerous dimensions that are used in various substantive fields. An outlier is a rare occurrence in a dataset that could be caused by a variety of factors. Finding and eliminating outliers from a data sample using simple univariate statistics such as standard deviation and interquartile range improve predictive modelling performance from a training dataset. The technique of finding outliers in a dataset is known as outlier detection. They defy the dataset's usual patterns. Outliers are predicted by a goal variable, but the rest of the points are typical. A clustering technique is used to find outliers, which skews the model's representation. Noise can take many forms, including incorrect data entry, mechanical faults, experimental failure, and natural diseases.

II. OUTLITER DECTECTION TYPES

- necessary.
- Unsupervised :There is no need to label the training data.
- Semi-supervised : an approach that combines supervised and unsupervised techniques

A. Applications

Detection of Fraud, Detecting the intrusion into the network Monitoring system activity and network performance, evaluating satellite photos, Keeping track of the passage of time, Data Leakage Prevention, Medical Report Diagnosis, Criminal Actions in E-Commerce, Video Surveillance, Anti-Terrorism, Pharmaceutical Research, recognising novelties, etc.

III. OBJECTIVE OF THIS PAPER

All Patterns that do not fit into a well-defined definition of normal behaviour are referred to as outliers. Anomalies, abnormalities, deviants, discordants, extreme data points, and other terms for outliers are all used to describe them. The phrases Outlier and Anomaly are interchangeable when it comes to revealing important information about a typical features. The following reasons make the technique extremely difficult: a) defining the line between normal and abnormal behaviour; b) model validation c) separating noise. b) cluster size distributed over space; c) stabilising the centroid point f) modifying unnormalized data into robust data;

***Corresponding author:**
E-mail:[1]rajaylakshmiravi7@gmail.com,[2]madhubalasivaji@gmail.com

[1]Research Scholar, Department of computer science,
Periyar University, Salem, Tamilnadu
[2]Research Supervisor , Department of computer science,
Periyar University, Salem, Tamilnadu

g) estimating fuzzy function point metric that gives more insight to imbalanced data in outliers using fuzziness

A. Organization of the Paper

The Outlier detection (also known as anomaly detection) is a crucial and difficult task in data mining for a variety of applications, including credit card fraud detection, intrusion detection, and image processing. They analyse the datasets in a meaningful and entertaining way. This study describes a method for detecting outliers that is both accurate and efficient. Outlier detection and prediction are inextricably linked. Outlier detection in its various forms is used in a variety of fields to identify intrusion, recognise anomalies, and so on. Detecting occurrences in credit card transactions that are fraudulent analysing system network traffic, diagnosing medical reports executing the legislation change, hastening the damage figuring out textual anomaly detection in the industrial sector figuring out how to detect textual irregularity in the industrial sector etc.,This study suggests that instead of deleting outliers, they should be worked with to fit within the boundary if necessary, as well as enhancing the estimation of various parameters that mask outliers during clustering.

An outlier in a dataset is an uncommon occurrence that can be caused by a multitude of circumstances. Simple univariate statistics such as standard deviation and interquartile range can help forecast model performance by identifying and removing outliers from a data sample. Compare typical and anomalous observations before declaring a datapoint as "outlier." Outlier is a term that refers to the intrinsic diversity of observation. Two types of outliers are possible. a) Excessive values (correct entry can not fit within cluster) b) Errors (wrong entry in the right place). The statistical analysis is hampered by missing values and null values, which lead to errors and disruptions. Error metrics is the most important phase in detecting the error from the prediction of evaluation.

B. Related Works

Unsupervised techniques of mixed data are studied in [1]. Applications, techniques in various domains are studied in [2]. A survey about outliers are studied in [3]. Cluster centroid stability using fuzzy is studied in [4]. Multivalued data streams using clustering is studied in [5].

C. Fuzzy clustering

Clustering is the process of identifying comparable groupings of data in a dataset. Clustering is a strategy for unsupervised learning that is subjective in nature. Soft clustering, often known as fuzzy clustering, is a type of clustering that has a tendency to be fuzzy. Simple implementation with dependable performance creates uncertainty in the model. In 1981, Jim Bezdek came up with the idea that for each member in the dataset, it discovers known variants using the membership [0,1] and prototype matrix.

The membership grade (or degree of membership) is calculated using Euclidean distance and statistical features of clusters.

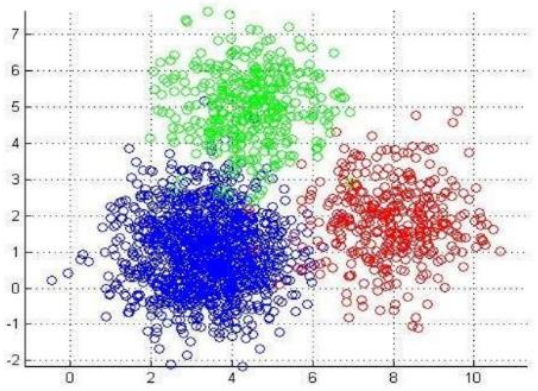

Fig. 1. Fuzzy Clustering

Consider a set S with a subset A(x) and elements defined as A(x) = 1 if membership degree is subset value 0, and A(x) > subset value if membership degree is > subset value. The fuzzy membership ranges from 0 to 1. Each data point is assigned membership, and the distance between the data point and the cluster centre is determined. The centroid centre, rather than the borders, has a high degree of membership for the data objects. A membership value of one corresponds to each data point.

D. Pseudocode of Fuzzy

1.Set up the membership matrix U.

2.Find the centre of the fuzzy cluster circle C.

3.Using Euclidean distance, calculate the distance between distinct circle centres.

4.Continue to calculate OF until it falls below the threshold.

5.Otherwise, To acquire the best value, fix a random point 'p' and repeat step 2 again.

E. Proposed Methodology

"There won't be any intruders among the outliers." They have a variety of effects on the regression line, including the appearance of normal predictor values, atypical predictor values following the line, and unusual predictor values not following the line. Outliers can be caused by a variety of factors, including incorrect entry, misreporting, sample error, and a rare but genuine value. a)Discarding b)Winsorizing c)Variable transformation d)Fit various models are some of the common methods for dealing with outliers. e)Dropping but not forgetting f)Non-parametric approaches.

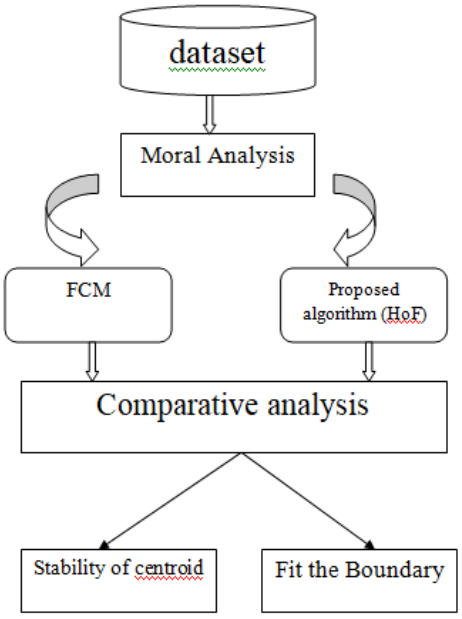

Fig. 2. The Framework of HoF Analysis

Outliers are points that are distinct from the rest of the data and provide useful information for the analysis. 1.data turns out to be skewed format 2.changes the overall statistical distribution of data in terms of mean,variance 3.Leads to obtain a bias in the accuracy level of the model. Figure 2 shows the framework of HoF.

F. System Model

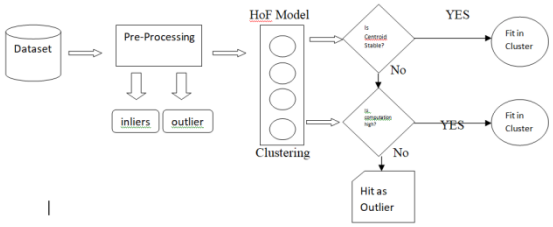

Fig. 3. Block Diagram of Proposed model (HoF)

G. Proposed Algorithm (HoF : Hit of Fit)

To create an objective function, run FCM. [HOF]

To find outliers, divide the data into tiny groups.

The remaining details (not in step 2)

Begin
Sum=0.5
Remove one point from each Pi point in the set.
Calculate HOFi using Pi

DOFi=(HOF-HOFi)/2 is the formula for calculating DOFi.
Sum =DOFi+Sum
Do Avg DOF=sum/n and return Pi.
Pi is equal to each point
Do If (DOFi> T) and return if point pi is an outlier.
Otherwise, come to a halt.

H. Experimental setup

Case analysis of Outlier Detection – centroid stability

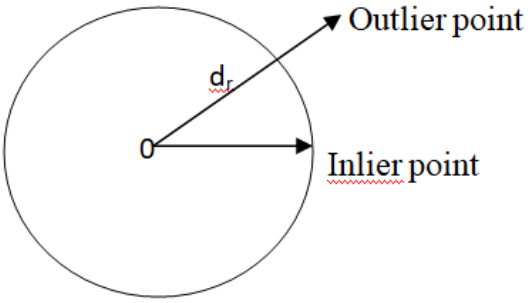

Fig.4 Stabilizing centroid point of the cluster

Consider a Dataset 'ds' with 'dpi' datapoints as inliers, 'dpo' datapoints as outliers and 'dr' as distance with reachability time. If reachability time is minimum, computational efficiency is high. From Fig. (3),

If dpi< threshold boundary, so called inliers

If dpo< threshold boundary, so called outliers.

TABLE ICENTROID STABILIY

Dataset	Samples	FCM (outliers detected)	HoF(outliers detected)	No of iterations
Stock	3000	yes	yes	250
Advertising	200	No	yes	250

TABLE 2
COMPUTATIONAL EFFICIENCY

Dataset	Samples	FCM (outliers detected)	HoF(outliers detected)	No of iterations
Stock	3000	yes	yes	250
Advertising	200	No	yes	250

From Table 1 and 2 it is clear that HoF algorithm works well in aspect of cluster centroid stability and high computational efficiency.

1.When data is normalised with a Z-score (-3 to +3), an IQR is calculated. (IQR) (Interquartile Range) (Interquartile Range) (Interquartile Range

2.Data is prepared using one of the following methods: a)Graphical Method b)Numerical Method

3.Analysis Techniques a)HighReverage Point b)Influential Observations c)Simple Linear Regression a)High Reverage Point b)Influential Observations c)Simple Linear Regression a)High Reverage4.Assessment Methods - Algorithms for fuzzy grouping and calculating centroid

IV. COMPARATIVE ANALYSIS

Before clustering consider the following points, i) determine the number of clusters to be stated. ii) Know how to optimize(k=3) and solve a problem before creating the method. iii) Understand why your data should be grouped into a certain number. Following table shows the comparative analysis.

TABLE 3
COMPARATIVE ANALYSIS OF DATASETS

Algorithm	Dataset	No of records	No of clusters	Result percentage
FCM	Advertising	200	3	44.06%
	Stock_data	3000	3	52.52%
Proposed algorithm(HoF)	Advertising	200	3	61.92%
	Stock_data	3000	3	69.48%

Thus from fig (4) and table 3 shows that proposed system works well for stock dataset in an efficient manner.

V. Conclusions

We obtain subjective knowledge and observation of the data through analysing. The proposed approach is compared with existing method by considering advertising and stock dataset to prove stability of the centroid and high computational efficiency.

ACKNOWLEDGMENT

I sincerely thank my research supervisor who encourages to complete my work a great success.

REFERENCES

[1] AtiqurRehman , Samir BrahimBelhaouari, "Unsupervised outlier detection in multidimensional data," Journal of Big data, 2021.
[2] Xiaodan Xu, Huawen Liu, Minghai Yao, "Recent Progress of Anomaly Detection," Hindawi, 2019.
[3] Hongzhi Wang, Mohamed Jaward Bah, Mohamed Hammad, "Progress in Outlier Detection Techniques: A Survey," IEEE, 2019.
[4] ErindBedalli, EneaMancellari, OzcanAsilkan, "A heterogeneous cluster ensemble model for improving the stability of fuzzy cluster analysis," Elsevier, 2016.
[5] Agnieszka Duraja, Piotr S.Szczepaniaka, "Outlier Detection in Data Streams – A Comparative study of selected methods", Elsevier, 2021.