Original Research Article

J. Computing & Int. Systems (2022) 299–302

SACRED HEART RESEARCH PUBLICATIONS

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in

ISSN: 2456-9496

AUTO-DETECTION, CHARACTERIZATION AND STATISTICAL EVALUATION METRICS MEASURES AND EFFICACY OF PERPETUAL OUTLIERS USING FUZZY

S Rajalakshmi[#1], P. Madhubala[#2]

Received on 13 NOV 2022, Accepted on 30 NOV 2022

Abstract — Due to advancement in the field of machine learning and data science, researchers show great interest over unusual behaviour. This paper tells not to remove outliers instead hit through to make it fit inside the boundary, if necessary and also to enhance the estimation of various factors that hides outliers optimal result during clustering. Outlier detection is an important research topic in the real time scenario. To face the overwhelming of uncertainity, factors makes the approach very challenging – a)to define the boundary between normal and anomalous behaviour (hitting and fitting) b) size of cluster distributed over the space c) stabilize the centroid point d) validation of models e) difficult to distinguish noise f) to impose unnormalized data into robust data g) estimate fuzzy function point metric that give more insight to imbalanced data using fuzzy function point analysis during outlier detection . The first stage aims to handle the reduction of noise. The second stage aims to cluster the data using fuzzy clustering. The third stage involves various algorithm by passing many iterations of k-means, fuzzy clustering and enhanced algorithm to stabilize centroid. Next, ends with the result analysis by finding the fuzzy-fp metric, fuzzy variable index and cluster utilisation factor. Final step validates the cluster goodness using silhouette, Pseudo-F-Statistic and Constellation method. Prons and Cons of this approach are listed in an effective manner. In large dataset, result shows a better performance that determines the factor by good performance score and add insight to give a balanced perspective on the data.

Keywords - Outliers, Centroid, Stability, Cluster, Fuzzy

I. INTRODUCTION

Outlier detection is the technique of identifying outliers within a dataset. They don't follow the typical patterns of the dataset. The remainder points are usual, but the outliers are anticipated by a target variable. Outliers are found using a clustering technique that skews the model's representation. Erroneous data entry, mechanical flaws, experimental failure, and natural diseases are all examples of noise.

An outlier is a rare occurrence in a dataset that can be caused by a variety of factors.To identify and remove outliers from a data sample, employ simple univariate statistics like standard deviation and interquartile range improve predictive modelling performance.

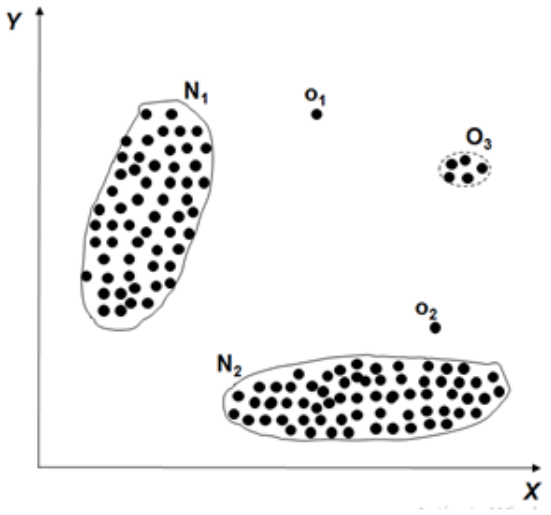

Fig. 1. Outliers

Before calling a datapoint as ‘outliers’ compare normal observation with abnormal observation. In short, the inherent of variability during observation is called outlier. There may be two set of outliers. a) extreme values(correct entry not fit in cluster) b) mistakes(wrong entry in the right place). Missing values and null values cause error and disturbs in the statistical analysis.

The degree of membership belongs from 0 to 1. In order to tolerate the uncertain and imprecise, it is a must to detect outliers. Section 2 includes the key challenges, Section 3 describes related works and stages of proposed work, Section 4 includes the Objectives and effect of outliers in the model, Section 5 describes the proposed fuzzy algorithm, Section 6 includes about the table of comparative analysis , Section 7 includes experiments and results discussion.

241

Picture retrieval using text, which looks for photographs based on one or even more criteria. Keywords identified by the user was the most popular image searching tool in the past. However, there are times when keywords cannot

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

[1]Research Scholar, Department of Computer Science, Periiyar University, Salem..

[2] Research Supervisor, Department of Computer Science, Periiyar University, Salem..

II. KEY CHALLENGES

- The boundary between normal datapoint and outlying datapoint is often not precise
- Identifying an exact normal region is quite challenging.
- The notion of outlier is different for different application domains
- Availability of labelled data contains noise.
- The process of starving interesting and fantastic knowledge in detecting Health care informatics to indicate disease outbreaks and instrumentation errors
- Normal labels available if misclassification cost is very high.

III. RELATED WORKS

Outlier techniques address the challenges, existing methods to evaluate outlier explanations[1].

For subsequent cluster prioritization, stability shows desirable surrogate for statistical measurement based on resampling and outlier boundary estimation[2].

A. Stages of the Proposed Paper Work

1.Data preprocessing

- Data quality is checked by Data Cleaning, Data Integration, Data Reduction, Data Transformation and Data Discretization.
- Noise is reduced by 'Binning' and the tools are Data Scrubbing, Data Auditing and Data Migration.
- Data Reduction is reducing the size by aggregation, sampling, dimensionality reduction and feature subset
  - Data is normalized by z-score, decimal scaling, min-max
  - Data Discretization is how data is sorted
  - Overfitting is too many variables used for training.

2.Data understanding

- clustering data into groups and stabilize the centroid
- fitting the boundary of the cluster
- fixing saturation point
- calculating utilisation factor c=̃λ/μ

3.Visualization techniques

- histogram
- boxplot
- constellation diagram

4.Statistical Evaluation metrics

- fuzzy-fp metric give more insight to imbalanced class.
- fuzzy variable index
- utilisation factor
- It also specifies the ordinal-type variables. Ordinal variables are measurements that may be ordered according to magnitude.
- For example, a survey question may require you to pick one of five possible choices: strongly disagree (5), disagree (4), neutral (3), agree (2), or strongly agree (1).
- Interval variables are ordinal, but ordinal variables are not necessarily interval. The original values of ordinal variables are replaced by their ranks. These ranks are then analyzed as if they were interval variables.

4.Cluster validation

- It measures cluster goodness be cluster separation and cluster cohesion
- silhouette (graphical analysis)
- pseudo f-statistic (statistical analysis) and constellation diagram.

IV. OBJECTIVES

The main aim of this paper is to detect perpetual outliers, stabilize centroid, estimate cluster boundary fitness and to evaluate statistical metrics over perpetual outliers

A. Effect of Outliers in the Model

1.The data turns out to be skewed format

2.It changes the overall statistical distribution of data in terms of mean,variance

3.Also, leads to obtain a bias in the accuracy level of the model.

All the input parameters are characterized using a factor called Cluster Utilisation Factor(CUF). CUF is evaluated by no of intensive clusters in 'inliers' and in 'outliers'. Cluster boundary is fitted using Cluster Utilization Factor(CUF). It is used for estimating the border of by

Size factor
Scaling factor
Geometric size in Magnitude

V. PROPOSED FUZZY ALGORITHM

a) Fuzzy clustering

Fuzzy clustering which is a soft clustering identifies the similarity measure over uncertainty. Clustering, and unsupervised model works well over huge dataset of unknown facts usinf fuzzy membership[0,1]. The degree of membership lies between 0 to 1. The distance is calculated by Euclidean distance from centroid to the cluster boundary.

b) Pseudocode of Fuzzy

The fuzzy pseudocode algorithm includes the following steps such as

1.Initialize membership matrix U
2.Calculate fuzzy cluster center C
3.Calculate OF until the value is below threshold.
4.Otherwise, Fix a random point 'p', to get optimum value–repeat from step 2.

c) *Cluster Boundary Fitness*

a. Clustering the outlier by fuzzy is the most challenging aspect in todays world. In order to eradicate these type of events, we proposed a new model fuzzy based outlier detection method to detect outliers. Initially, Data is pre-processed to reduce the noise and focused on predictions to reveal highly desirable task. No of clusters is

b. determined for the iterations. Distance is evaluated using Euclidean measure. Fuzzy objective function is measured based on the threshold value. The Cluster validation is determined by the optimal no of clusters.

d) *Proposed Framework Model*

Outliers are data points that are far from the original data that gives useful information for the analysis. It leads to some negative statistical measures in many applications and the performance, computational complexity, time complexity and parameter estimation vary from time to time. Fig.2 shows the framework model of proposed system

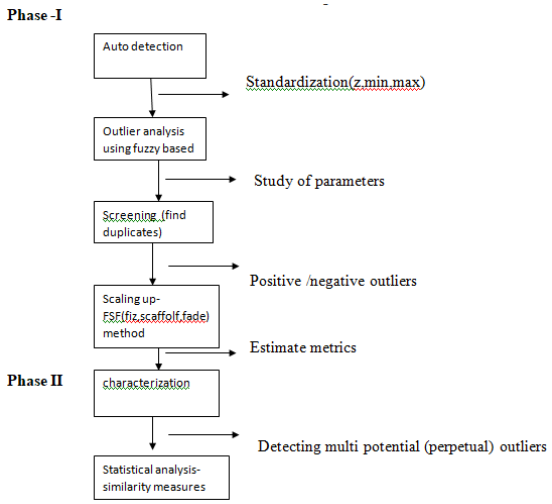

Fig. 2. The Proposed Framework

The proposed model includes two phase: Phase I and Phase II.

In Phase I, Auto-detection of perpetual outliers is identified by z-score and max-mix normalization. The parameters are studied to estimate the statistical metrics.

In Phase II, Characterization is done to achieve accuracy and efficacy over similarity measure of data points. Finally, perpetual outliers is identified for centroid stability, cluster boundary fitness and statistical metrics evaluation accuracy.

Proposed Algorithm (CBF : Cluster Boundary Fitness)

Execute FCM to generate objective function [OF]
Assume X and Y be inner and outer boundary
Begin
Sum=(X+Y)/2
For each iteration in the set Do
Remove a point from outer boundary cluster
Calculate OFi with inner boundary cluster
Calculate DOFi =(OF-OFi)/2
Sum=Sum + DOFi
Return average of two successive cluster boundaries
End do
Avg DOF=sum/n
For each point Pi
Do
If (DOFi > T)
consider Xc as CBF into Yc as an CUF
and return Xc
Else stop
End do
End

VI. TABLE I

CLUSTER BOUNDARY FITNESS

Dataset	Samples	FCM (outliers detected)	Proposed Model (CBF)	No of iterations
Advertising	200	FALSE	TRUE	249
Airport noise monitoring	120	TRUE	TRUE	249
Auto	398	FALSE	FALSE	249
Ch10Ex11	1000	TRUE	TRUE	249
College	778	FALSE	NEUTRAL	249
Credit	400	FALSE	FALSE	249
Data_Uer _Modeling	259	FALSE	FALSE	249
Directory of food stamp centers	18	TRUE	TRUE	249

TABLE 2

STATISTICAL EVALUATION METRICS

Dataset	Samples	'z' test	'chi-square' test	't' test
Advertising	200	>1	<0	0.076
Airport noise monitoring	120	<0	<0	NA
Auto	398	>0	NA	NA
Ch10Ex11	1000	NA	0.46	3.5
College	778	NA	1.88	NA
Credit	400	<7	>3	1.42
Data_Uer _Modeling	259	<0	>2	NA
Directory of food stamp centers	18	<0	<0	NA

From Table 1 it is noticed that CBF algorithm suits well for detecting cluster boundary. Table 2 shows the statistical measures and efficacy over outliers of 't' test.

VII. COMPARATIVE ANALYSIS

By comparing the number of iterations from FCM to Proposed Model, our methodology works well to identify perpetual ouliers. The iterations increased from FCM to Proposed CBF (Cluster Boundary Fitness) model is

experimented by R programming 4.2.0. Statistical evaluation includes 't' test, 'chi-square' test and 'z' test. Among all the three test 't' test well suits for our proposed model.

VIII. CONCLUSIONS

By analyzing, we gather subjective knowledge to process the learning data, perform iterative calculations by optimizing position of the centroids and defined number of iterations achieved. This method enhances accuracy, efficacy over similarity measure and performance efficiency that assessed over some UCI repository datasets. Experimental results and statistical evaluation measure shows effectiveness of Centroid Stability and efficacy of Cluster Boundary Fitness to detect perpetual outliers.

ACKNOWLEDGMENT

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

REFERENCES

[1] Panjei E, Gruenwald L, Leal E, et.al, "A Survey on Outlier Explanations," Th VLDB Journal, 2022. https://doi.org/10.1007/s00778-021-00721-1

[2] Tianmou Liu, Han Yu, Rachael Hageman Blair, "Out-of-bag stability estimation for k-means Clustering," Statistical Analysis and Data Mining: The ASA Data Science Journal, 2022. https://doi.org/10.1002/sam.11593.

[3] Tobias Ziolkowski, Agnes Koschmider, Peer Kroger, Colin Devey, "Outlier quantification for multibeam data," Informatik Spektrum, 2022. https://doi.org/10.1007/s00287-022-01469-w

[4] Michal Bechny, Johannes Himmelbauer, "Unsupervised approach for online outlier detection in industrial process data," Procedia, 2022. https://doi.org/10.1016/j.procs.2022.01.224

[5] Andre Q, "Outlier Exclusion Procedures must be blind to the researcher's hypothesis," Journal of Experimental Psychology: General, 2022. https://doi.org/10.1037/xge0001069

[6] Ehsan Jolous Jamshidi, Mohamad Anuar Kamaruddin, "Detecting Outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature," Ecological Informatics, 2022. https://doi.org/10.1016/j.ecoinf.2022.101672

[7] Mehhdi Jabbari Nooghabi, Mehrdad Naderi, "Stress-strength reliability Inference for the Pareto distribution with Outliers," Journal of Computational and Applied Mathematics, 2022. https://doi.org/10.1016/j.cam.2021.113911

[8] Atiq ur Rehman, Samir Brahim Belhaouari, "Unsupervised Outlier detection in multidimensional data," Journal of Big data, Springer, 2021.