Data Analytics and Artificial Intelligence
Vol: 2(5), 2022
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/data-analytics-and-artificial-intelligence/

Detection of Perpetual Outliers by Scaffolding Centroid Stability and Fitness over Cluster Boundary using Fuzzy

[*]**S.Rajalakshmi, P.Madhubala, M.Saseekala**
Periyar University, Salem, TamilNadu, India.
*Corresponding author Email:rajaylakshmiravi7@gmail.com

Abstract. Detection of outliers is increasing exponentially every year. This paper includes new Outlier detection method to detect the perpetual outliers. The challenging factor is identifying the unsharp boundary that cannot define data point lies inside or outside. The first stage includes scaffolding to stabilize centroid data points. The second stage includes to fix the cluster boundary over fitness. The third stage includes comparative analysis. To achieve an optimal solution, our method gives good result and stability over the cluster border. The overlapping factor in n-dimensional space is also rectified using scaffold tri-vatiate dimensional space. Keywords: Outliers, Centroid, Stability, Cluster, Fuzzy.

1. Introduction

The challenges in outlier detection focus data model and the application used. The common techniques that is used to deal with outliers are a) deleting observations that are too far from real dataset b) transforming and binning values-variable transformation c) Imputing missing values.

The impact of outliers on dataset are

- Outliers change the results of statistical and data analysis
- It reduces power of statistical tests
- It increases error variance that decreases normality
- Include values and assumptions from other statistical models such as regression and ANOVA.

2. Outlier Analysis

An outlier is an observation in a random sample of a population that is abnormally distant from other values. The most popular methods of outlier detection are

- Z-score (or) Extreme Value Analysis
- Probabilistic and Statistical Modeling
- Proximity Based Models
- Linear Regression models
- Information theory Models
- High Dimensional outlier detection methods

There are two categories of outliers. Univariate outliers are one-dimensional space and Multivariate outliers are n-dimensional space. Outlier comes in a variety of flavours depending on the surroundings. They are

I. Point outliers
II. Contextual Outliers
III. Collective outliers.

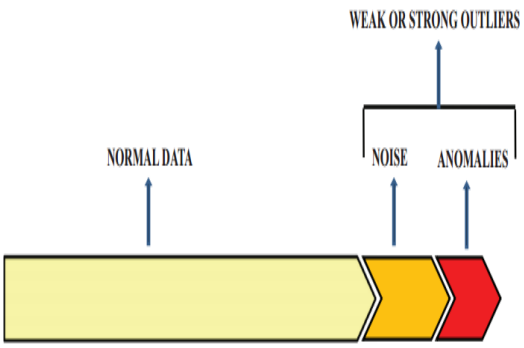

FIGURE 1. Outlier

3. Clustering

Cluster analysis is the process of gathering related objects together through unsupervised learning. There are many different kinds of clustering techniques, including partitioning techniques (k-means, k-medoids), grid-based techniques (Wavecluster, clique, sting), hierarchical techniques (divisive, dendogram, and agglomerative), model-based techniques, and density-based techniques (denclue, optics, dbscan). It is widely used in many applications, including data analysis, pattern recognition, market research, and image processing using software programmes like R programming, Weka tool, Python, and SPSS.Clustering enables business sectors to divide their clientele into groups according to common trends [1]. Additionally, it aids in the classification of documents, spatial data analysis, housing planning, insurance prediction, earthquake analysis, and market research in the economy. The many types of properties, arbitrary shapes, noise, outliers, scalability of data objects, constraint-based grouping, and interpretability are all covered.

4. Fuzzy Clustering

It is the method of clustering used mainly for pattern recognition, where one data point belongs to two or more clusters. It is developed by Dunn in 1973 and further improved by Bezdek in 1981. Fuzzy C-means, which is an extension of k-means clustering and finds soft clusters, is also known as fuzzy k-means. FCM is not a full member of the specific cluster. However, the data point has a slight advantage over the cluster of points. To determine the density of membership of data in the cluster, fuzzy coefficient centroids are calculated [3].The n-dimensional vector space is used to determine the distance measurement. The distance from the point to the centroid is directly proportional to the affinity value.

5. Related works

Outlier detection techniques in financial oriented sectors are studied from [1]. Centroid gravity with coefficient method is studied from [2]. Distance basedOutlier detection in high dimensional dataset is studied from [3].

6. Organization of the paper

There are many case stuies in Bioinformatics – Gene Detection, Stock Marketing and Credit Card Fradulent. In the scenario of fraudulent of normal transaction, high recall with false positives is measured for model's performance as f2 score.Centroid methods: It gives better result than other algorithms but computationally expensive

- COG centre of Gravity - It is a point on vertical line which it divided in to two halves and overlapping area is counted once. It is calculated by
- COG = a (x,μ(x) d(x))
- COS Centre of Sum-Computationally less expensive where overlapping area is counted twice
- COA Centre of Area - It is the simplest method, where subregions are mentioned.

We focused on stability of the centroid by scaffolding the datapoints. From Fig.2, 3 random datapoints are chosen for stability of the centroid.

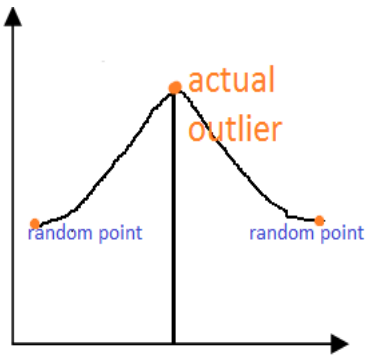

FIGURE 2. Scaffolding the centroid data point

The distance is measured from scaffolding data points to the cluster boundary. Euclidean distance is calculated by Eq. (1)

$$J(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}(\mu ij)^{m}\,\|xi-vj\|^{2} \qquad (1)$$

$||xi- vj||$ is Euclidean distance between i^{th}data and j^{th}cluster center.

μij -membership matrix

'm' is the fuzziness index m € [1, ∞].

'c' represents the number of cluster center

Proposed Methodology:

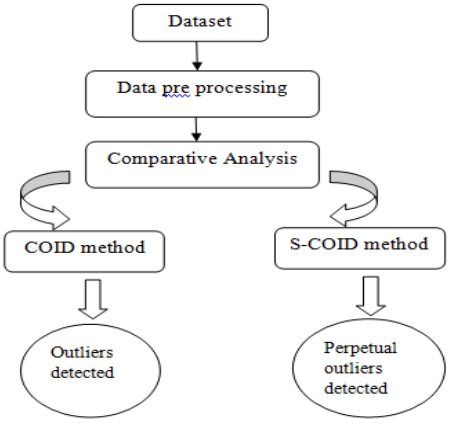

FIGURE 3. S-COID proposed method

Proposed Algorithm:

Step 1: Generate FCM
Step 2(a): Make a pass over the data points to identify 3 random points for scaffolding - centroid stability
Step 2(b): Make another pass over the data points to identify cluster boundary fitness
Step 3: Calculate the distance from scaffolding centre to the cluster boundary
Step 4: Combine step1 and step2 to generate Outlier threshold value
Step 5: If the scaffold data point is greater than the outlier threshold value declare as perpetual outlier
Step 6: Detect the outliers and separate it during another pass.
Step 7:Repeat until threshold value remains normal

7. Outlier Evaluation Metrics

Clustering tendency, non-uniform distribution is solved by Hopkins test as null hypothesis (H0), and alternate hypothesis that are used to analyse clusters (Ha). The quality of the clustering, which is determined by metrics such the intrinsic measure-silhoutte and extrinsic measure. Number of ideal clusters (k), which is determined by a data-driven technique as an empirical and elbow method as well as a domain knowledge approach. Statistical approach of Clusters are represented by values of the gap statistic.

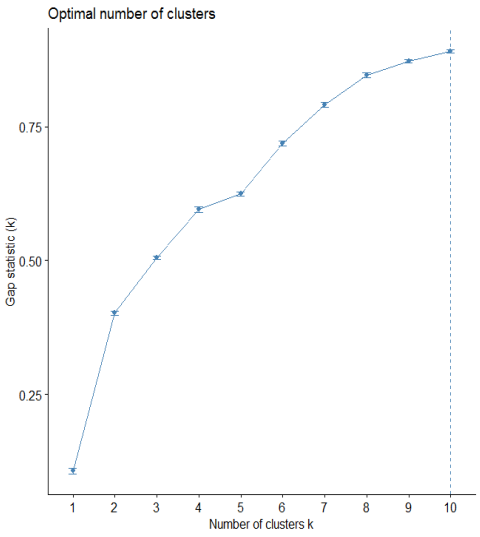

FIGURE 4. No of clusters using Gapstatistic

The purpose of the cluster validity indices is to compare two sets of clusters or clustering techniques in terms of connectivity and compactness. It establishes whether a cluster is caused by noise. Internal, external, and relative cluster validity are the metrics used to assess cluster validity. A) Single linkage distance; B) Complete linkage distance; C) Average linkage distance; D) Centroid linkage distance.

Outlier evaluation metrics lack this level of accuracy, however projected labels compared to training labels provide justification for perpetual outliers. Total number of true positives over the sum of true positives and false positives is referred to as precision. Total true positives over both true positives and false negatives are referred to as recall. The confusion matrix is used to better understand the quantiles, which are referred to as f1 scores.

8. Comparative Analysis

The S-COID proposed method is compared with COID method using R Script Programming 4.2.0. The precision and the recall value remains same as nearest to 74 % . Statistical evaluation calculation and comparative analysis is shown below.

TABLE 1. Statistical Evaluation Metrics

Dataset	Samples	'z' test	'chi-square' test	't' test
Dycd-contractors	12337	NA	2.22	>1
Heart	304	NA	>0	NA
Income1	31	0.2	2.22	0.33
Income	31	>1	NA	
Jobs_Proximity_Index	219985	<0	<0	NA
LinkedInTopSkills	51	<0	>2	NA
NYC_Jobs	4191	>1	=1	NA
OCC_Output	703	NA	NA	>1
Stock_data	3000	<1	0.55	1.31

Table 1 shows the statistical measures of proposed method and efficacy of outliers in't' test.

TABLE 2. Comparative Analysis

Dataset	Samples	COID	S-COID
Dycd-contractors	12337	TRUE	TRUE
Heart	304	TRUE	TRUE
Income1	31	TRUE	TRUE
Income	31	TRUE	TRUE
Jobs_Proximity_Index	219985	FALSE	FALSE
LinkedInTopSkills	51	FALSE	FALSE
NYC_Jobs	4191	NA	NA
OCC_Output	703	FALSE	FALSE
Stock_data	3000	TRUE	TRUE

Table 2 shows the comparative analysis of proposed method and COID [3]. The precision and recall remains same for both the algorithm using R programming 4.2.0. But the time efficacy of S-COID over COID shows better enhancement over time.

9. Conclusions

Thus the proposed S-COID method gives an excellent solution in identifying perpetual outliers. It also achieves its goal of providing a structured and generic solution. Thus our method gives an exciting results over uncertainity towards mathematical model. Though the high dimensional dataset remains unpredictable, this paper shows a good behavioural output and understanding over perpetual outliers. It works well on many applications for better understanding of outlier detection techniques for future enhancement also.

Acknowledgment: I sincerely thank my research supervisor who encourages to complete my work a great success. Also I thank my friend for her gradual support in clarify my doubts.

References

[1]. YallaVenkateswarlu, K. Baskar, AnupongWongchai, VenkateshGauri Shankar, Christian Paolo Martel Carranza, José Luis Arias Gonzáles, A. R. MuraliDharan, "An Efficient Outlier Detection with Deep Learning-Based Financial Crisis Prediction Model in Big Data Environment", Computational Intelligence and Neuroscience, vol. 2022, Article ID 4948947, 10 pages, 2022. https://doi.org/10.1155/2022/4948947

[2]. Bin Li, Xingping Bai, Jun Li, Lirong Wang, "Outlier Detection of Gravity Dam Deformation Monitoring Data Based on the Multiple Local Outlier Coefficient Method", Mathematical Problems in Engineering, vol. 2022, Article ID 7157844, 10 pages, 2022.https://doi.org/10.1155/2022/7157844

[3]. Ankit Kumar, Abhishek Kumar, AliKashif Bashir Mamoon Rashid, V. D. Ambeth Kumar, "Distance Based Pattern Driven Mining for Outlier Detection in High Dimensional Big Dataset", ACM Transactions on Management Information Systems. 10.1145/3469891, 2022 Vol 13 (1) pp. 1-17

[4]. Pedro Aguiar, António Cunha, MatusBakon, Antonio M. Ruiz-Armenteros, Joaquim J. Sousa, Multivariate Outlier Detection in Postprocessing of Multi-temporal PS-InSAR Results using Deep Learning, Procedia Computer Science, Volume 181, 2021, Pages 1146-1153, ISSN 1877-0509,

[5]. https://doi.org/10.1016/j.procs.2021.01.326.

[6]. https://www.analyticsindiamag.com/most-popular-clustering-algorithms-used-in-machine-learning/