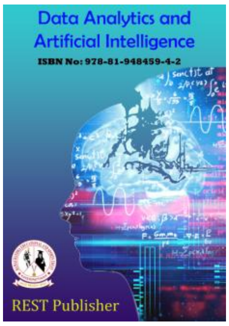
Data Analytics and Artificial Intelligence

Vol: 3(2), 2023

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/

DOI: https://doi.org/10.46632/daai/3/2/4


Data Analytics and Artificial Intelligence
ISBN No: 978-81-948459-4-2
REST Publisher

Certain Investigation on Perpetualistic Fuzzy Outlier Data for Efficiency Evaluation of Centroid Stability with Cluster Boundary Fitness

* **S. Rajalakshmi, P. Madhubala**

Periyar University, Salem, Tamilnadu, India.

*Corresponding Author Email: rajaylakshmiravi7@gmail.com

Abstract. *This paper aims to investigate certain factors that hide outliers in two dimensions such as boundary partitioning and space angular parameters. In this proposed algorithm, boundary representation of clusters, the data points that lie on the cluster boundary is stored geometrically as coordinate values such as i_bound (inliers) and o_bound(outliers). Outliers that present in dataset are investigated by boundary fitness over centroid stability. In this paper we focus to examine whether the data point lie on the boundary is treated as inliers or outliers. Several iterations are manipulated to fix the outlier point deeply. Using fuzzy clustering, data points are clustered and boundary is fixed. If the space occupied by the cluster varies for every iteration, the distance from inlier to outlier between the boundaries is calculated. After calculation, if the data point is below the threshold value, it is treated as outlier. Our proposed method shows efficiency over evaluation metrics of outlier detection performance.*

Keywords: *Outliers, Cluster Boundary, Centroid Stability, Angular Parameter, Fuzzy*

1. INTRODUCTION

Data mining is a non-trivial extraction of knowledge used to identify hidden patterns, correlation patterns, predictive information and decision making information. There are many technical approaches such as classification, clustering, association rules, prediction and estimation. Among these techniques, Clustering is an unsupervised learning. It is the categorization of items into sets of related objects (clusters). Usually, the objects are described as feature vectors (also called attributes). A vector is used to describe item x if one has n properties (x1... xn).

The reasons why clustering is necessary in data mining are explained by the following points: Scalability: To handle high dimensional datasets, we require clustering techniques that are more extremely scalable. Capability to handle various properties: Algorithms should be able to process any type of data, including categorical, binary, and interval-based (numerical) data. Finding clusters with a shape attribute: The clustering method should be able to find quality clusters with any shape defined. High dimensionality: In addition to handling high-dimensional data, the clustering algorithm should be able to handle space over boundary fitness. Capability to handle noisy data

2. MOTIVATION AND OBJECTIVE OF THE PAPER

The main aim of this research work is to enhance the estimation of numerous factors that hide outliers to produce the best clustering results. Though many Outlier Detection methods are existing, it lacks in computational complexity, accuracy level in investigating outliers around boundary and thus makes the research very interesting.

Some ***factors*** that makes the approach very motivating and challenging.

- ✓ Defining the boundary between normal and anomalous behaviour
- ✓ Validation of models
- ✓ Difficult to distinguish noise
- ✓ Cluster size distributed over space
- ✓ Stabilizing the centroid point
- ✓ Modifying normalized data into robust data
- ✓ Estimating fuzzy function point metric that gives more insight to imbalanced data in outliers
- ✓ Making predictions difficult

is outliers poblemmatic? why?

Yes, because of the following challenges.

- ✓ 1.The data turns to skewed format
- ✓ 2.Changes in statistical results
- ✓ 3. Clustering leads to bias in the accuracy level of the model.
- ✓ 4.Unusual behaviour of patterns
- ✓ 5. Affect regression line by unusual predictor value

So, while preparing dataset for machine learning model, it is important to detect the outliers.

3. RELATED WORKS

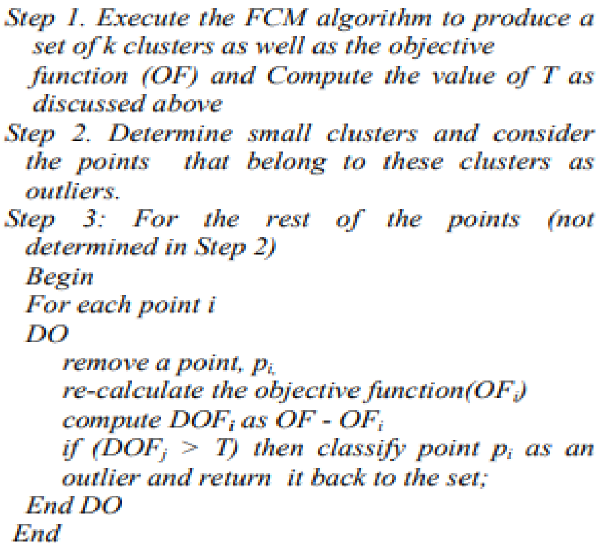
Two clustering algorithms, such as centroid based k-means and representative object based FCM are compared based on number of datapoints and number of clusters are studied from [1]. From [2] Fuzzy c-means Clustering is studied how the segmentation of clustering takes place in presence of noise. Stability of cluster is studied by 2 factors such as number of different cluster of previous timestamps [3]. Also stability score on depends subsequence score of all cluster members is studied from [3]. RFCM algorithm suitable for noisy data and unequal clusters is studied from [4]. Semi-supervised method in order to improve detection precision and to reduce false alarm rate is studied from [5]. Also Cluster center point, lower approximation with fuzzy boundary using FCM clustering is studied from [5]. Selection of initial prototype based clustering method for cluster centroids and how to speed up the computation time I studied from [6].

4. OUTLIER DETECTION

Outlier is a rare occurrence, where the patterns not fit to normal behaviour. The other terms of outliers are abnormalities, anomalies, discordants, deviants or extreme data points . Its types are point, collective, contextual. A clustering method that skews the model's representation is used to find outliers. Examples of noise include incorrect data entry, mechanical issues, unsuccessful experiments, and natural diseases. A dataset outlier is an unusual occurrence that can be brought on by a number of different causes. Simple univariate statistics like standard deviation and interquartile range can be used to locate and eliminate outliers from a data sample and enhance the efficacy of predictive modelling. Applications are fraud detection, intrusion detection, event detection, image processing, health care informatics, enforcing law amendment, industrial sector, under surveillance, medical diagnosis, detection of business failures, textual anomaly, predictive maintenance and so on.

5. FUZZY CLUSTERING ALGORITHM

It is a soft 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, is an extension of k-means clustering. To determine the density of membership of data in the cluster, fuzzy coefficient centroids are calculated. Assigning objects to groups of related objects is known as clustering (clusters). Typically, the objects are described as vectors of features (also called attributes).



```
Step 1. Execute the FCM algorithm to produce a
   set of k clusters as well as the objective
   function (OF) and Compute the value of T as
   discussed above
Step 2. Determine small clusters and consider
   the points that belong to these clusters as
   outliers.
Step 3: For the rest of the points (not
   determined in Step 2)
   Begin
   For each point i
   DO
       remove a point, $p_i$,
       re-calculate the objective function($OF_i$)
       compute $DOF_i$ as OF - $OF_i$
       if ($DOF_j$ > T) then classify point $p_i$ as an
       outlier and return it back to the set;
   End DO
  End
```

FIGURE 1. Existing Algorithm

Proposed Algorithm: The proposed algorithm is stated below.
1. to create an objective function, run FCM.
2. to find outliers, divide the data into tiny groups.
3. The remaining details (not in step 2)

```
Begin
i_bound=0
o_bound=1
Sum=0.5
Remove one point from each Pi point in the set.
Calculate distance from i_bound and o_bound using angular parameter Pi(ϕ,θ)
DOFi=(i_bound+o_bound)/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.
```

To rectify the problem refer fig.2,

- ✓ first cluster the data,
- ✓ second calculate the centroid
- ✓ for stabilization use Proposed model (i_bound, o_bound)
- ✓ third calculate OF(objective function)
- ✓ Finally, if datapoint is below threshold declare as outlier.

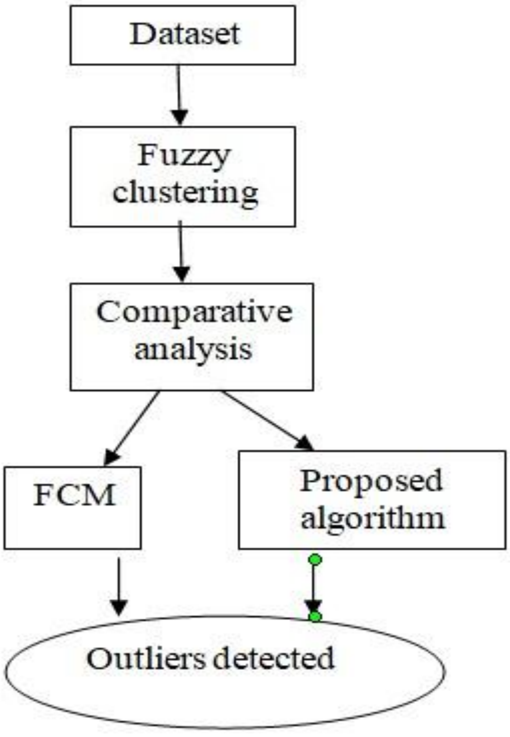


FIGURE 2. Proposed work model

Boundary Fitness to stabilize Centroid: These iterative clustering methods use the proximity of a data point to the cluster's centroid to determine how similar two objects are. It is crucial to have prior knowledge about the dataset in these models since the number of clusters needed at the conclusion must be specified beforehand. In order to locate the local optimum, these models run iteratively as shown in figure 2.

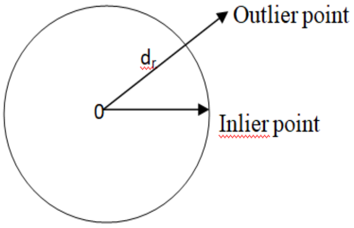


FIGURE 3. Investigation distance between two points

b)Angular Parameter to fix Centroid Point: Stability of a cluster is defined by *i_bound* and *o_bound.* The parameter determines the distance between the two spaces. The n-dimensional vector space is used to determine the distance measurement. **Distance Measurements** Consequently, object x is described as a vector if one has n properties (x1,..,xn) is shown in figure 3. These common distance functions can be defined for two p-dimensional data items, i_bound = (xi1,xi2,...,xip) and o_bound = (xj1,xj2,...,xjp) by *Euclidean Distance Function* as shown in equation(1)

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2} \quad (1)$$

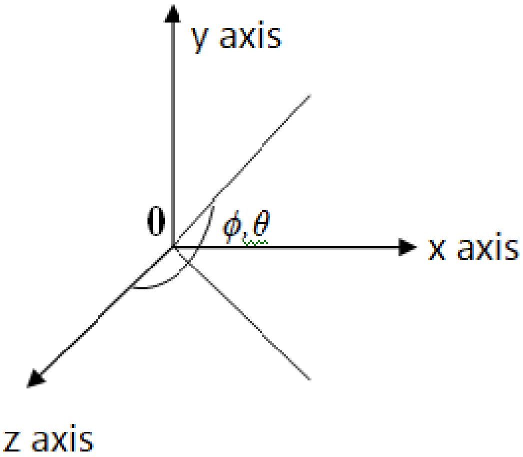


FIGURE 4. Angular parameter touching boundary

6. COMPARATIVE ANALYSIS

Cluster validity is a collection of methods for identifying the clusters that best suit a set of given datasets. A cluster validity index verifies the result. From Table 1 and 2, comparative analysis of FCM and Proposed algorithm is clearly mentioned. Number of ideal clusters (k), is established by elbow method of gap statistic. Internal, external, and relative cluster validity are the metrics used to assess cluster validity. Outlier evaluation metrics lack relative level of accuracy; however projected labels compared to training labels provide justification for perpetual outliers in table 1.

Precision: Total number of true positives over the sum of true positives and false positives is referred to as precision. ***Recall:*** Total true positives over both true positives and false negatives are referred to as recall.

TABLE-1 Comparative analysis

Dataset	Centroid Stability	FCM	Proposed algorithm	Error rate greater
Iris	YES	TRUE	TRUE	NO
Stock data	YES	TRUE	TRUE	YES
Advertising	YES	FALSE	FALSE	NO

The quality of the clustering is determined by metrics such the intrinsic measure-silhoutte and extrinsic measure as shown in table 2.

TABLE 1. Evaluation metrics analysis-silhoutte

n	**i_bound**	**o_bound**	**Max (i_b, o_b)**	**Angular parameter Pi=i_b-o_b/ max(i_b,o_b)**
0	2	8	8	8-2/8=0.75
2	0	6	6	6-0/6=1.00
4	2	4	4	4-2/4=0.50
6	2	4	4	4-2/4=0.50
10	2	8	8	8-2/8=0.75

7. CONCLUSION

The proposed approach shows better result while compared with existing method. It also enhances accuracy of detecting perpetual outliers. The estimation over centroid stabilization achieves high computational efficiency. Experimental results shows effectiveness of detecting perpetual outliers by poorly fit of boundary during clustering by investigating the stabilization of centroid point. I sincerely thank my research supervisor who encourages completing my work a great success.

REFERENCES

[1]. Soumi Ghosh, Sanjay Kumar Dubey, “Comparative Analysis of K-Means and Fuzzy C-Means Algorithms,” International Journal of Advanced Computer Science and Applications, Vol.4, No.4, 2013.

[2]. Royna Daisy. V, Nirmala. S, “Stability Integrated Fuzzy C-Means Segmentation for Spatial incorporated automation of No of clusters,” Indian Academy of Sciences, Sadhana, 2018.

[3]. Klassen, G., Tatusch, M. & Conrad, S. Cluster-based stability evaluation in time series data sets. Appl Intell (2022). https://doi.org/10.1007/s10489-022-04231-7

[4]. Salar Askari, “Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development”, Expert Systems with Applications, Volume 165, 2021, 113856, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.113856.

[5]. Xue, Zhenxia & Shang, Youlin & Feng, Aifen. “Semi-supervised outlier detection based on fuzzy rough C-means clustering”, Mathematics and Computers in Simulation. 80. 1911-1921. 10.1016/j.matcom.2010.02.007.

[6]. Cebeci, Zeynel & Cebeci, Cagatay, “A Fast Algorithm to Initialize Cluster Centroids in Fuzzy Clustering Applications”, Information Switzerland 11. 446. 10.3390/info11090446.