IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Auto-Detecting Perpetual Outliers Using Efficient Modified Fuzzy Clustering Approach

S.Rajalakshmi
Research Scholar
Department of Computer Science
Government Arts College for Women-Krishnagiri

P.Madhubala
Research Supervisor
Department of Computer Science
Don Bosco College – Dharmapuri

Abstract:
To create some outstanding new ideas of innovation during Pandemic, this paper reveals an augmented study of modified robust fuzzy clustering approach to detect the unusual outliers. Outlier Detection, an unsupervised learning used to detect complex dataset to identify perpetual outliers using membership function to tolerate the uncertainty. The first stage aims to estimate auto-detection of noise reduction, incorrect data points and so on as perpetual outliers. The second stage aims to determine the characterization of utilized input parameters of unending outliers using FCM (Fuzzy C-means Clustering) method to find FOF(Fuzzy Objective Function). This method enhances accuracy, efficacy over similarity measure and performance efficiency that assessed over some UCI repository datasets. Experimental results and statistical evaluation shows effectiveness of detecting perpetual outliers.

Keywords: Outliers, fuzzy, clustering, Fitness, Perpetual, FOF, Characterization

1. Introduction:

During Pandemic, the situation is worst and uncertain to predict the future events. Outlier detection is also an important and complex task due to its uncertainty intolerance. Outlier known as anomaly identifies the extreme points from the dataset [1]. Fuzzy clustering is the suitable method for handling the uncertainty. Due to technology development, rare events happen due to technology advancement. Data is pre-processed to reduce the noise and focused on predictions to reveal highly desirable task. No of clusters is 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 ofclusters. All the input parameters are characterized using a threshold value. It is evaluated by no of intensive clusters in *'inliers'* and in *'outliers'*. The degree of membership belongs from 0 to 1. Accuracy is very close to cluster input parameters of detected prototypes. During Characterization, a random learning of false data point and dominating capability of influence structure is fixed.

2. Literature Survey

Outliers which may treated as error results in underestimation of uncertainty tolerance. Outlier analysis is studied from charu.C.Aggarwal[1][3][7]. The various types of anomaly detection and characteristics is studied from chandola, Banerjee and Kumar[2]. Clustering techniques are identified by aggarwal,chandan and reddy[3]. Outliers identity is studied from Hawkins[4][9]. How to detect the outliers in real time application using fuzzy clustering is studied from Rajalakshmi and Madhubala[5].R programming by Yanchang Zhao[9].Fuzzy set is studied from Bezdek[11] and klir yaun[10].

The Prediction model is used to calculate residuals of estimating various waves of fuzzy outliers. From the study, the non-adherence of protocol *stemmed* up by the characterization of extreme datapoints. The number of clusters indicate efficient generalization over learning of fuzzy objective function by less computational effort. It uses membership value, cardinality value, and cluster validation indices.

A modified approach using feature based indexing for the labeled patterns with high membership function is considered for generalization [7]. A generalized fuzzy index method and novel constraint function of membership is understand by the study [8]. A detailed study of estimating ERR and ERR for z-score is demonstrated in [14]. How the psych and z-curve package is evaluated in R-script to study the outlier fitness as well from [17][18][19].

3. About the software packages used:

R tool (r 4.0.2 version) is used to analyse the outliers using fuzzy clustering. "psych" package is installed for factor analyzing the Perpetual outlier fitness. William Revelle[17][18], clearly estimates ERR((Expected Replicability Rate), EDR(Expected Discovery Rate) and ODR(Observed Discovery Rate) using the packages *psych* and *z-curve* .Confirmatory factor analysis is estimated by using fuzzy clustering. *iclust(),* item cluster analysis used to partition space of the fitness rather than space of the variable. It explains an implementation of z-curves and method for estimating replicability rates (ERR) and expected discovery rates(EDR) on finding fitness, metrics etc

Algorithm
Phase I:
Step 1: Initialize no of clusters 'k'
Step2: Compute fuzzy membership matrix 'U'
Step3: calculate fuzzy centers 'Vj' until j is minimum
Step4: Estimate the centroid of each cluster center 'c'
Step5: Update the minimum distance to each cluster using Euclidean distance
Step6:Auto-detect the centers to identify farthest neighbor
Step7:If the object has no neighbor, define it is outlier

Algorithm
Phase II:
Step 1:Initial the fuzzy variables
Step2:Calculate the sum of all the objects which has similarities
Step3:Fix a random point between two similar clusters (say for eg. Between 0.5 and 0.7, 0.6 as random intensity values)
Step 3a: Compare the random object with threshold value.
Step 3b: If it is equal to threshold, move to next neighbour cluster
Step 3c: Repeat step 3a until the value is below threshold
Step4:At the end, extreme points from the old cluster to is used to detect the perpetual outlier of new cluster
Step5: Depending on threshold value, outlier is identified.
Step6:The objective function is obtained from new cluster is called FOF (Fuzzy objective function)

4. Experimental Results

RStudio 4.0.2 is used for estimating statistical analysis. Packages used are "psych" and "zcurve", where zcurve developed by Frantisek Bartos on 27,September -2020 to detect the accuracy of outliers[17][18][19]. By installing *zcurve* package, we used density method to find the increase number of iterations and decrease number of criterion for the dataset stock_data. Estimation of ERR(Expected Replicability Rate) and EDR(Expected Discovery Rate) is shown below.

Table 1: Estimation of z-curve 2.0 density algorithm

Estimation of z-curve 2.0 density algorithm		
Dataset	ERR	EDR
iris	NA	NA
Stock_data	0.6135274	0.5064392
Advertising	NA	NA

Using *zcurve* package, we used density method for the version 1 to increase the number of iterations for the dataset stock_data. Estimation of ERR(Expected Replicability Rate) and EDR(Expected Discovery Rate) is shown below.

Table 2: Estimation of z-curve to increase the number of iterations

Estimation of z-curve to increase the number of iterations		
Dataset	ERR	EDR
iris	NA	NA
Stock_data	0.6479701	0.5046194
Advertising	NA	NA

Using *zcurve* package, we furnished the details regarding total number of starting fits and the means of mixture components. Estimation of ERR(Expected Replicability Rate) and EDR(Expected Discovery Rate) is shown below.

Table 3: Estimation of z-curve to increase the number of starting fits and change in means of mixture component

Estimation of z-curve to increase the number of starting fits and change in means of mixture components		
Dataset	ERR	EDR
iris	NA	NA
Stock_data	0.5880518	0.4578420
Advertising	NA	NA

In the fourth table, we comprised the z-curve fitness to simulate z-statistics. The estimation of ERR(Expected Replicability Rate), EDR(Expected Discovery Rate) and ODR(Observed Discovery Rate) is shown below.

Table 4: Estimation of z-curve to simulate z-statistics and z-fitness for stock_data dataset

Estimation of z-curve to simulate z-statistics and z-fitness for stock_data dataset			
z-score	Estimation Range (0.64 – 5.51)	Percentage	CI-(Confidence Interval)
ERR	0.82	95%	0.73,0.89
EDR	0.55	95%	0.24,0.89
ODR	0.85	95%	0.80,0.89

In the fifth table, we calculated the power components of z-score for the value of alpha from 0 to 1.(0.05 to 0.999)

Table 5: Estimation of z-curve power components

Estimation of z-curve power components							
Power to z-score	I	II	III	IV	V	VI	VII
Range (0-1)	0.05	0.20	0.40	0.60	0.80	0.974	0.999
Reult	0.000197359	1.11451417	1.706298343	2.213272223	2.801581787	3.903097682	5.050194967

Table 6: Estimating the model to detect outliers using EM and confidence interval

Model 1: EM via EM					
z-score	Estimation of EM	l.CI	u.CI	Iterations	Outliers detected
ERR	0.620	0.501	0.746	45 +118	Q= -60.61
EDR	0.391	0.082	0.693		CI[-70.48, -47.79]

Table 7: Estimating the model to detect outliers using KD2 and confidence interval

Model 1: KD2 via EM			
z-score	Estimation of KD2	Iterations	RMSE
ERR	0.614	47	0.11
EDR	0.506		

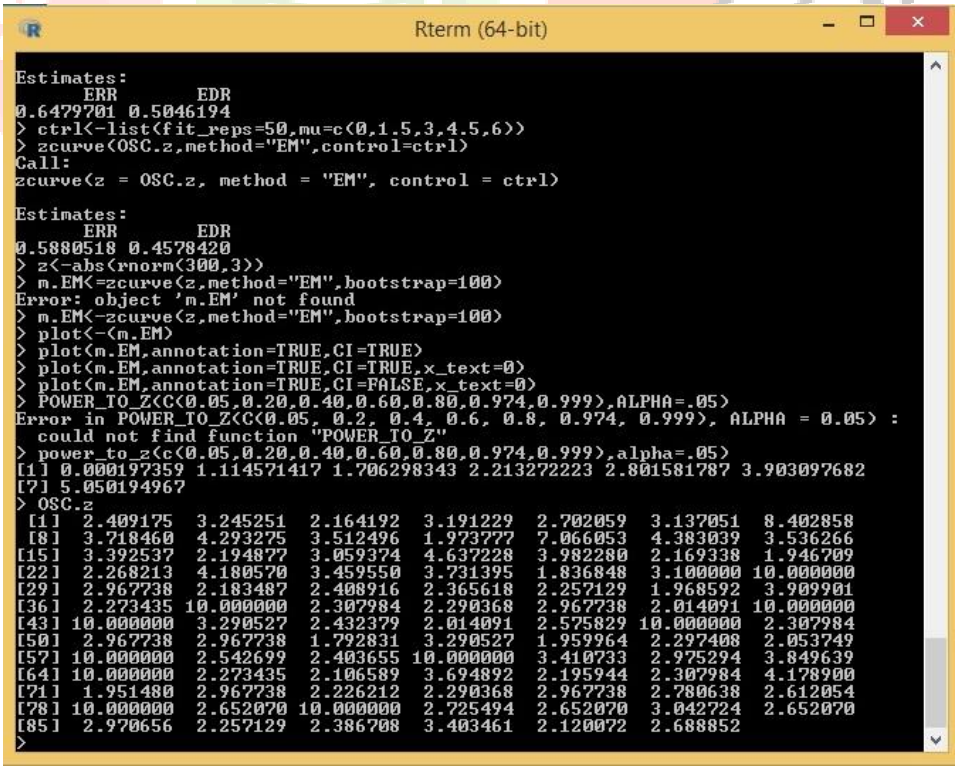

Figure 1: No of clusters after removal of noise in stock__data dataset

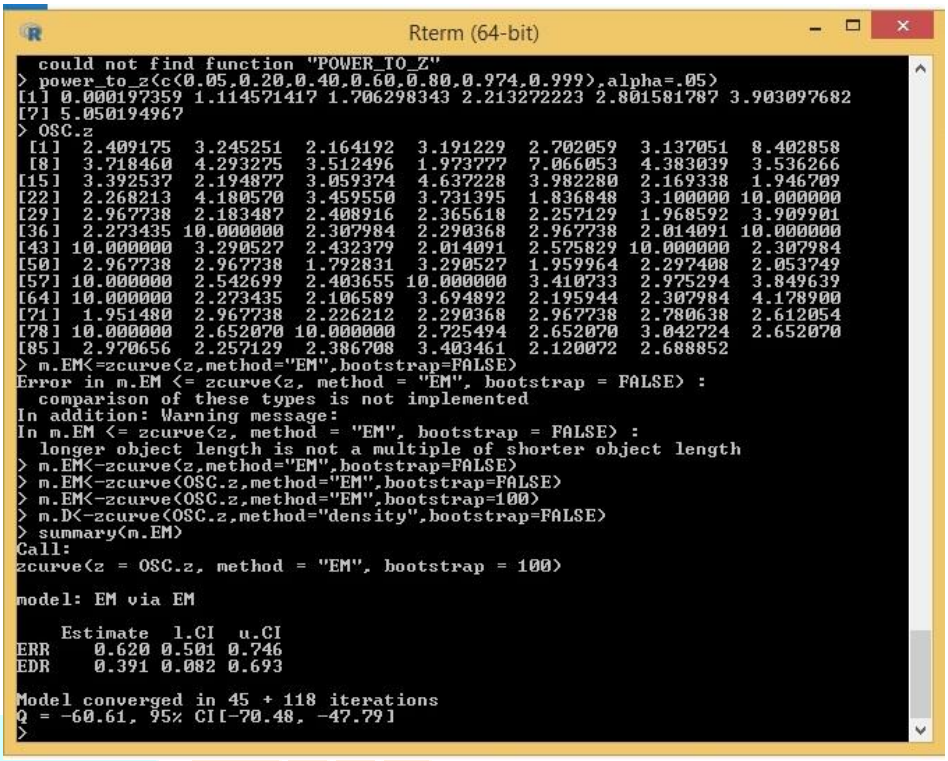

Figure 2: The Estimation of Using Expectation Method (EM) via Expectation Method (EM), for lower Confidential interval (l.CI) and upper Confidential interval (u.CI) for replicable iteration is as follows.

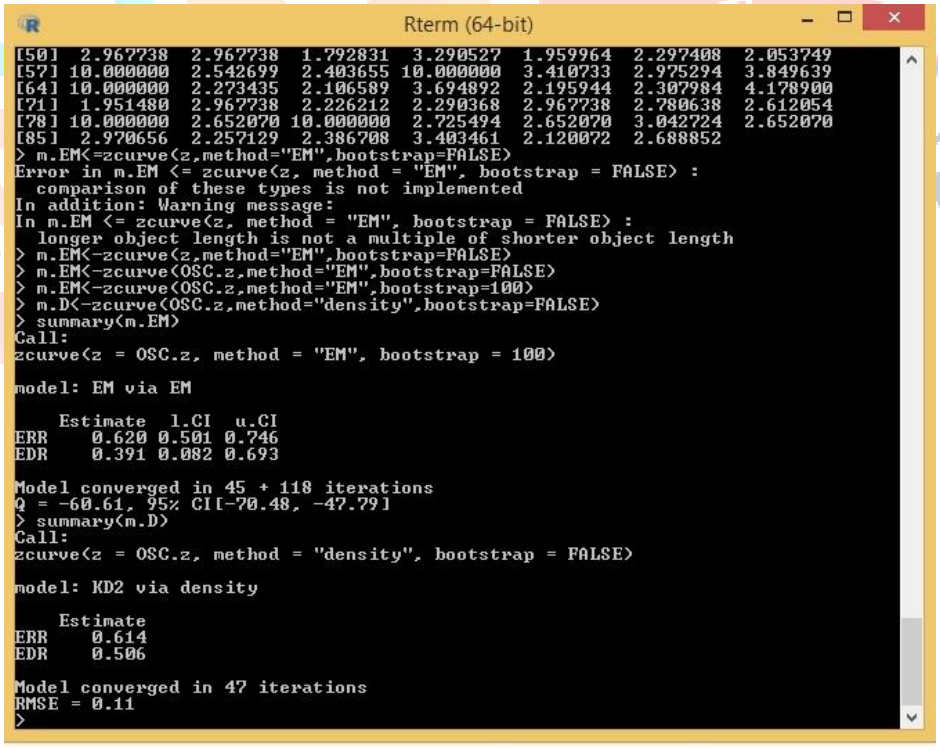

Figure 3: The Estimation of KD2 via Expectation Method (EM), for replicable iteration

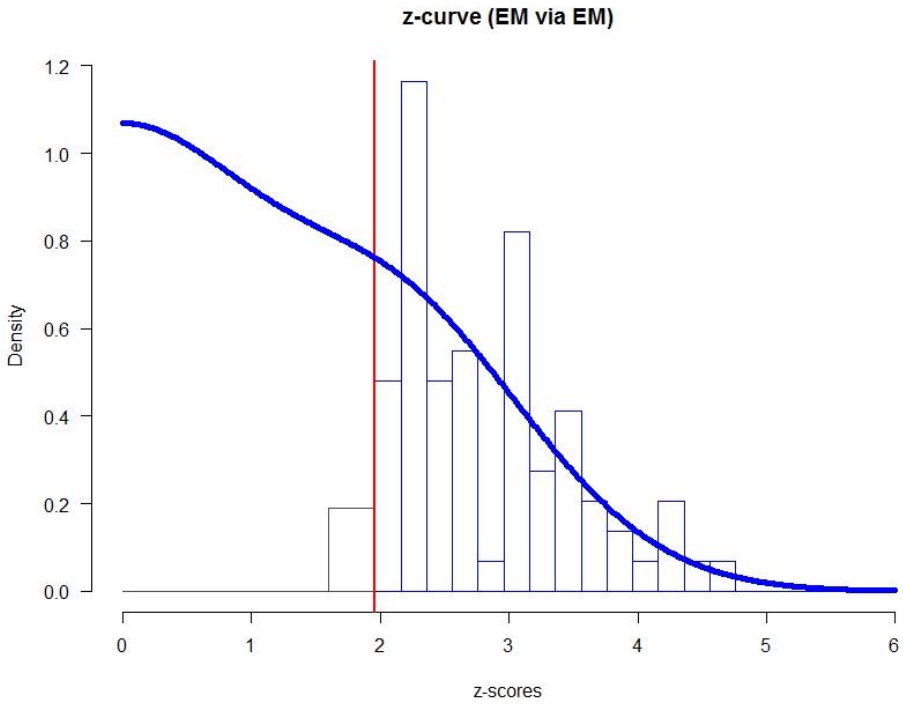

Figure 4: z-curve(EM Via EM)

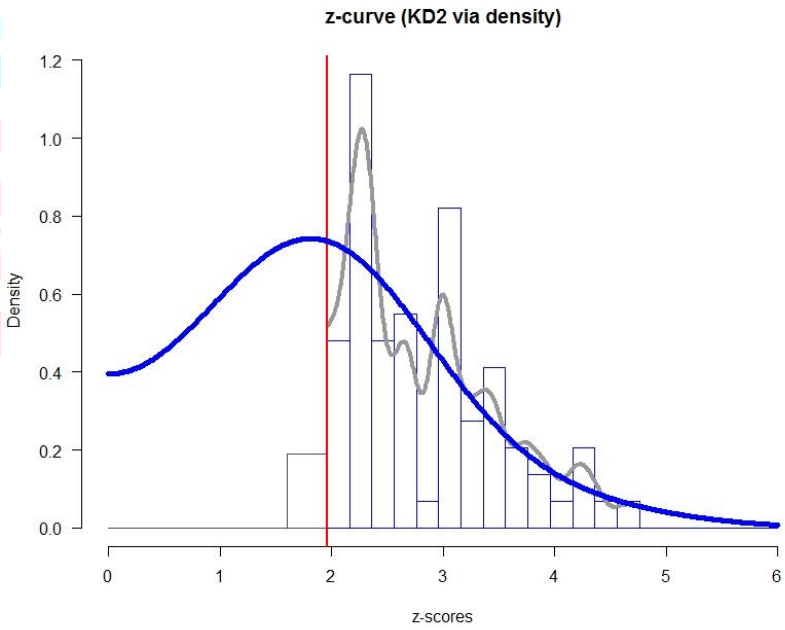

Figure 5: z-curve(KD2 Via DENSITY)

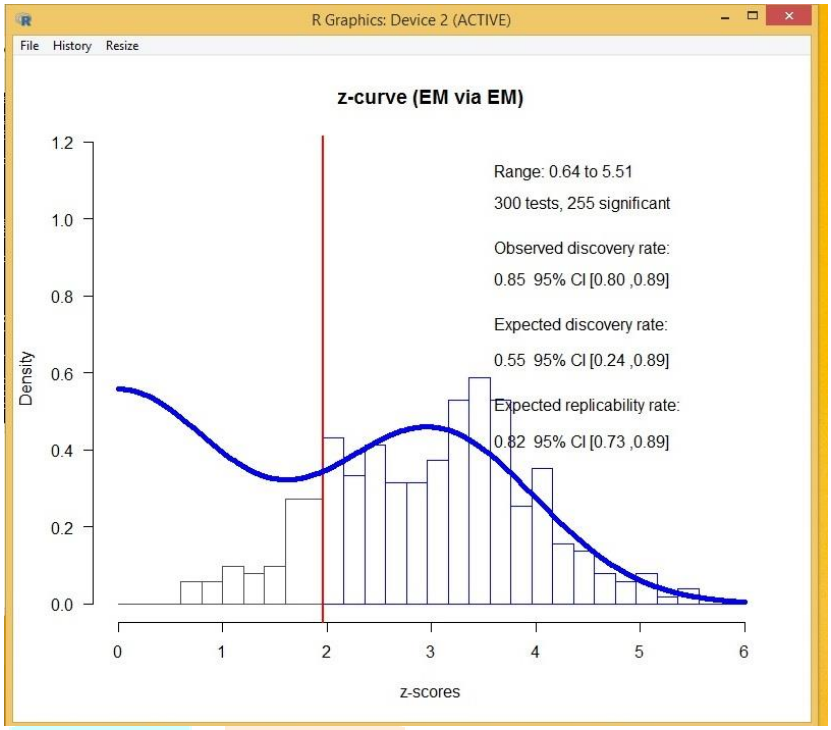

Figure 6: Using Expectation Method (EM),

From the figure (1) - (5) z-curve is estimated for many iterations of stock_dat the model converged at the rate of 0.09215481 with (25+66) iterations and (24+22) iterations with 92% detection of outliers.

FOF ~ ERR and EDR

5. Conclusion

Towards the contemporary world of pandemic, researchers are making very close and accuracy to the machine understandable efforts rather than human. Using utilization factor, the number of observation comes under the study is more efficient for the outcome of the outliers. Estimating is based on replicability rates of statistical testing such as t-test, F-test, chi-square test, Z-test, ERR and EDR. It simulates the robustness of outliers that undergoes uncertainty as significant to the margin. FOF(Fuzzy Objective Function) determines the utilization factor by fitting random points of clusters and detecting the perpetual outliers (disabilities) to fade the replicability fitness over the outlierst to normalize. The test results shows a complete consent to measure the values of outliers using fuzzy clustering approach.

References

[1].Charu C.Aggarwal, “Outlier Analysis”, Springer,2013.
[2].V.Chandola, A.Banerjee and V.Kumar, “Anomaly Detection: A survey”, ACM
[3].Charu c.Aggarwal, Chandan, K.Reddy, “Data clustering”, CRC Press ,2014.
[4]. Hawkins,”Identification of outliers”,1980.
[5]. S.Rajalakshmi,P.Madhubala,” Outlier Detection: A research and Modified method using Fuzzy Clustering”,IJITEE,ISSN:2278-3075,vol9,Issue 3s,Jan(2020).
[6].F.Hoppner,F.Klawonn,R.Kruse,”Fuzzy cluster Analysis:Methods for classification, Data Analysis and Image.
[7].Charu C.Aggarwal,Manish Gupta,Jing Gao,”Outlier Detection for Temporal Data: A Survey”,IEEE, Jan-2014.
[8].Irad Ben Gal,”Outlier Detection”,Data Mining and Knowledge discovery Handbook,kluwer Academic Publishers,2005.
[9].Yanchang zhao,”R and Data Mining: Examples and Case studies”,May 29,2012.
[10].Klir Yuan,”Fuzzy set and fuzzy logic-Theory and Applications”- BOOK
[11]. J D Harris, J C Bezdek, “Fuzzy Partition and relation - An axiomatic basis for clustering”, Fuzzy sets and systems, Elsevier, 1978.

[12]. Francesco Marcelloni, Feature selection based on a modified fuzzy c-means algorithm with supervision, vol 151, pages 201-226, May 2003.

[13].Lin Zhu et.al, Generalized fuzzy c-means clustering algorithm with improved fuzzy partitions, IEEE, June 2009.

[14].Bartoš, F., & Schimmack, U. (2020, January 10). Z-Curve.2.0: Estimating Replication Rates and Discovery Rates. https://doi.org/10.31234/osf.io/urgtn

[15].Bartoš, F., & Schimmack, U. (2020, January 10). Z-Curve. : An R package for fitting Z-Curves", R package version 1.0.6.

[16]. William Revelle, Department of Psychology Northwestern University, How To: Install R and the psych package, September 8, 2020

[17]. William Revelle, Department of Psychology Northwestern University, An introduction to the psych package: Part I: data entry and data description, September 4, 2020

[18]. William Revelle, Department of Psychology Northwestern University, An introduction to the psych package: Part II Scale construction and psychometrics, August 12, 2020

[19]. Package'zcurve' September 27, 2020.