P- ISSN 1991-8941 E-ISSN 2706-6703
2025,(19), (01):260 – 268

*Journal of University of Anbar for Pure Science (JUAPS)*

Open Access

# New Modified Technique to Identify Outlier Values by Using Fuzzy Clustering

**Saja Mohammed Sakran . Wafaa S. Hasanain***

Department of Mathematics, College of Science, Mustansiriyah University, Baghdad, Iraq
**Email:** w.s.hasanain@uomustansiriyah.edu.iq or w.s.hasanain@gmail.com

**ARTICLE INFO**

Received: 08/06/2024
Accepted: 21/07/2024
Available online: 21/06/2025

**DOI: 10.37652/juaps.2024.150715.1273**

**Keywords:**
***Fuzzy c-means, Possibilistic c-means, Standardization, Objective function, Simulation.***



**ABSTRACT**

Outliers within a dataset are data points that substantially differ from the rest of the data. These atypical data points can be attributed to several factors, such as errors in measurement, issues with the data input, and natural variations in the data. Managing outliers is essential to ensure the integrity of statistical analyses and avoid obtaining misleading results.

These outliers can be observed at very high or very low points and can exert a notable effect on statistical measures, such as mean and variance. Many diagnostic techniques focus on influencing centroids and distances between clusters to detect these abnormal points.

In this study, fuzzy cluster techniques are employed to identify outliers within a dataset. An alternative technique is utilized to detect outliers via standardization by using fuzzy cluster techniques. The performance of the proposed method is compared with that of other approaches through simulation.

## Introduction

Cluster analysis is useful and efficient for classifying large amounts of data, so it is suitable for further processing data groups. It can also be employed to manage datasets, identify outliers, and determine which variables should be combined initially and which should be considered separately. Researchers use hard and fuzzy cluster analysis methods for different goals and purposes. Zadeh [1], [2] introduced fuzzy sets in 1965 and defined an object that allows the modeling of inaccurate models mathematically. Since then, the method has been used widely to manage ambiguous data and simulate human inference procedures.

Hard cluster partitioning combined with either a hard isodata technique or a hard c-means algorithm was implemented in the initial version of the fuzzifier factor [3].

Then, in 1981, fuzzy C-means (FCM) [4], a well-known fuzzy algorithm based on partitioning, was introduced.

By determining the distances between each data point and the cluster centers, FCM allocates a fuzzy membership degree to each one. Dunn [5] presented a fuzzy version of this algorithm to handle data that belong to many clusters at the same extent. Gustafson and Kessel's clustering technique (GK-1979) [6] uses an adaptive distance norm as an extension of the basic FCM algorithm, which employs Euclidean distance to identify clusters with various geometrical shapes. Ohashi [7] attempted to adjust for noise by modifying the FCM method to obtain robustness against some outliers. Dave [8], [9] proposed the idea of noise clustering by splitting the objective function into two terms. The first term corresponds to the objective function for probabilistic clustering, and a noise cluster is used to represent the second term. Bandemer et al. [10] organized data analysis into four progressively difficult levels to detect

*Corresponding author: Department of Mathematics, College of Science, Mustansiriyah University, Baghdad, Iraq
**ORCID:https:// https://orcid.org/ 0000-0003-4486-8525 ,**
**Tel: +964 7712220141**
**Email: w.s.hasanain@uomustansiriyah.edu.iq**

the nature of data and treat them. To solve the FCM noise problem, Krishnapuram and Keller (1993), [11], [12] presented the possibilistic C-means (PCM) clustering approach. The strategy differs from previous clustering algorithms in that the membership values can be regarded as probability levels of the points belonging to the classes, and the resulting data partition can be viewed as a possibilistic partition. Pal et al. [13], [14] (1997) constructed the fuzzy PCM method, which is commonly known as the mixed C-means algorithm that combines the characteristics of FCM and PCM. Yang and Wul [15] (2006) developed the possibilistic clustering algorithm, which began a new algorithmic line aimed at improving FCM and PCM techniques. To make their proposed algorithm robust to noise and outliers, the authors suggested that the membership that results from it should be treated as an exponential function. Wu et al. [16] (2010) introduced the unsupervised possibilistic fuzzy clustering algorithm as a means to overcome the coincident cluster problem in PCM and the noise sensitivity issue in FCM.

**Methods and Materials in Fuzzy Clustering**

Clustering techniques are an effective tool for minimizing dimensions and identifying outliers. Through the use of several distance metrics, clustering allows the original large dataset to be divided into many groups of comparable objects on the basis of similarity difference features. Then, each group can be replaced by the most representative object that is located in the cluster center [17].

Clustering and partitioning algorithms aim to divide a dataset of *n* objects with *p* variables or features into *k* cluster subsets of data or clusters. A data point that represents or specifies a cluster is referred to as a prototype in the context of clustering.

**FCM Method**

Fuzzy clustering, sometimes referred to as soft clustering or soft k-means, allows data points to be included in several groupings. A membership grade, which indicates which cluster the data points belong to, is assigned to each point [18].

The FCM clustering algorithm was initially examined by Dunn [5] in 1973, and Bezdek [4], [19] generalized it in 1974. Unlike in the K-means method, in FCM, each data object is a member within each cluster, and membership degrees vary between 0 and 1. By minimizing the weighted within-group sum of squared errors, the iterative clustering technique divides the dataset into *k* partitions. Moreover, the FCM clustering algorithm is an unsupervised method that permits a single data observation to be a part of many clusters. With this feature, it can be helpful in identifying outliers by recognizing data points that do not firmly belong to any cluster. The following text shows a robust algorithm that uses FCM to identify outlier values.

The objective function of FCM is

$$obj_{FCM}(Xf: Uf, Kf) = \sum_{i=1}^{n}\sum_{j=1}^{k}(uf_{ij})^m \left\|xf_j - cf_i\right\|^2. \quad (1)$$

The fuzzier, mf, in the objective function specifies how fuzzy the clustering result is, and $1 \le m \le \infty$. Usually, two are chosen. Large values of *m* produce fuzzy clusters, and small values produce tough clusters. If *m* = 1, FCM turns into a hard algorithm and uses K-means to obtain the same results.

FCM needs to meet the following constraints:

i) $\sum_{j=1}^{k} Uf_{ij} = 1 \quad ; \ 1 \le i \le n;$

ii) $0 < \sum_{i=1}^{n} Uf_{ij} < n \quad ; \ 1 \le j \le k; \qquad (2)$

iii) The following update equations are used to minimize the FCM objective function.

iv) $$uf_{ij} = \frac{1}{\sum_{r=1}^{k}\left[\frac{\left\|xf_j - cf_i\right\|}{\left\|xf_j - cf_r\right\|}\right]^{\frac{1}{(m-1)}}} \quad ; \ \text{if } d_{ij} > 0; \quad (3)$$

v) For i = 1… k ; j=1… n;

vi)

$$cf_i = \frac{\sum_{j=1}^{n}(uf_{ij})^m x_j}{\sum_{j=1}^{n}(uf_{ij})^m} \ ; \forall\, i = 1, \ldots\ldots, k; \quad (4)$$

$$df_{ij} = [(xf_j - cf_i)'\,(xf_j - cf_i)]^{\frac{1}{2}}. \quad (5)$$

For I = 1, 2… k ; j=1, 2… n,

$$\left\|Uf^i - Uf^{i-1}\right\| < \varepsilon f. \quad (6)$$

Data points that fall apart in any cluster can be categorized as outliers once the algorithm has converged. Compared with hard clustering approaches, this methodology allows a more flexible and robust identification of outliers.

**PCM**

PCM clustering was created to overcome some of the limitations of the FCM algorithm. As a solution to the noise problem in FCM, Krishnapuram and Keller (1993) [11], [12] presented the PCM clustering technique.

The data partition in this technique can be interpreted as a possibilistic partition, and the interpretation of the membership values can be viewed as the degrees of the possibility that the points belong to the classes to determine the parameter. However, PCM must be run on the fuzzy clustering results of FCM. PCM's performance is highly dependent on initialization and frequently decreases because of the simultaneous clustering problem, even though it solves the noise sensitivity issue of FCM (Filippone et al., 2007) [20]. The objective function of PCM 1 is

$$obj_{PCM}(Xf, Tf, Cf) = \sum_{i=1}^{n} tf_{ij}{}^{\eta}\, d^2\,(xf_i, cf_j) + \sum_{j=1}^{k} \Omega f_j \sum_{i=1}^{n} \left(1 - tf_{ij}\right)^{\eta}. \quad (7)$$

The first component in the objective function above minimizes the weighted distances, and the second term suppresses the trivial solution (Timm et al., 2004) [21].

An alternate objective function for PCM was proposed by Krishnapuram and Keller (Krishnapuram & Keller) [12]. The objective function of PCM 2 is

$$obj_{PCM}(Xf, Tf, Cf) = \sum_{i=1}^{n} tf_{ij}{}^{\eta f}\, d^2\left(xf_i, cf_j\right) + \sum_{j=1}^{k} \Omega f_j \sum_{i=1}^{n} tf_{ij}{}^{\eta f}\, log tf_{ij}{}^{\eta f} - tf_{ij}{}^{\eta f}, \quad (8)$$

Where

$$\Omega f_j = \frac{K \sum_{i=1}^{n} uf_{ij}{}^{m} d^2\,(xfi\ , cfj)}{\sum_{i=1}^{n} uf_{ij}{}^{m}}, \quad (9)$$

where

Xf = $\{xf_1, xf_2, \ldots\ldots, xf_n\} \subseteq R^p$ is the dataset for *n* objects in *p*-dimensional data space *R*;

Cf = $\{cf_1, cf_2, \ldots\ldots cf_n\} \subseteq R^n$ is the protoype matrix of the clusters;

Uf = $\{\mu f_{ij}\}$ is the matrix for a fuzzy partition of (Xf);

Tf = $\{tf_{ij}\}$ is the matrix for a possibilistic partition of (Xf); and

$d^2\left(xf_i, cf_j\right)$ is the squared Euclidean distance between object xj and cluster prototype (Cf).

$$d^2(xfi, cfj) = \|xfi - cfj\|^2 = (xfi - cfj)^T (xfi - cfj) \quad (10)$$

(mf) is the fuzzifier to specify the amount of fuzziness for the clustering; $1 \le mf \le \infty$ is usually chosen as 2.

Ηf is the typicality exponent to specify the amount of typicality for the clustering; $1 \le \eta f \le \infty$ is usually chosen as 2.

FPCM must satisfy the following constraints:

$\sum_{j=1}^{k} uf_{ij} = 1 \quad ; 1 \le i \le n,$

$\sum_{i=1}^{n} tf_{ij} = 1 \quad ; 1 \le j \le k.$

The membership degrees can be defined as typicality values that measure the degree to which a data object is for a particular cluster independent of all other clusters because PCM membership computation is possibilistic. The typicality degree update equation, which is obtained from the PCM objective function, is the same as that of FCM.

$$tf_{ij} = \left[1 + \left(\frac{d^2(xfi, cfj)}{\Omega f_j}\right)^{1/(m-1)}\right]^{-1} \quad (11)$$

For $1 \le i \le n$ , $1 \le j \le k$

The update equation for cluster prototypes is the same as those of FCM.

$$cf_j = \frac{\sum_{i=1}^{n} tf_{ij}{}^{m} xfi}{\sum_{i=1}^{n} tf_{ij}{}^{m}} \quad ; \quad 1 \le j \le k \quad (12)$$

Outliers are data points that do not fit well into any cluster and can be identified using PCM clustering.

**Proposed Method**

Accurately diagnosing outliers in datasets improves analytical model accuracy and data quality. Effective approaches include fuzzy clustering techniques, such as PCM and FCM. However, variations in measurement among observations in a dataset might affect the efficacy of these techniques. In this study, we propose a new strategy on the basis of the concept of standardization to enhance outlier diagnosis by utilizing FCM and PCM. We also examine the possible effects of this strategy on the results.

The process of standardizing data includes placing data within a common range so that the variables are compared using the equation

$$Zf_i = \frac{Xf_i - \mu f_i}{\sigma f_i}, \quad (13)$$

where $\mu f_i$ and $\sigma f_i$ are the mean and standard deviation of $Xf_i$, respectively.

This strategy recommends standardizing the data before applying fuzzy clustering techniques (FCM and PCM) to increase the precision of outlier diagnosis. This strategy's main effects are as follows:

- Increase the consistency of the data: The effects of data that are abnormally large or small must be minimized to ensure correct aggregation.
- Minimize the effects of disparate measurements: Every variable must fall within the same range to reduce the effects of measurement variations.
- Increase the precision of the diagnosis: The distinctions between normal and outlier values must be made obvious to increase the precision of the diagnosis.

**Criteria for Evaluating the Performance of Methods Used in Diagnosing Outlier Values**

1- Objective function: the lowest value of the objective function should be obtained
2- Skill of understanding and interpreting the results by diagnostic accuracy of outlier values
3- Computational efficiency is measured as the time or iterations needed to complete the algorithm
4- Small groups are identified and classified as outliers when the data points located in these little clusters have a distance Df (xf, cf) value equal to twice the mean of the distance value or greater or less than it as follows:

Df (xf, cf) $\geq 2$ ($\overline{\text{Df}}$; mean of distance) or Df (xf, cf) $\leq 2$ ($\overline{\text{Df}}$; mean of distance). (14)

**Simulation and Model Estimation Results**

Monte Carlo experiments are conducted using the MATLAB R22 program to assess the behavior and statistics of the methods employed for detecting outliers, and the methods are compared with the proposed strategy. Then, the effectiveness of these methods in the diagnosis process is assessed.

In particular, a dataset is generated via simulation with multiple sample sizes (nf = 10, 25, 50, and 100), different dimensions (pf = 2 and 3), and various numbers of clusters (kf = 2, 3, and 4). Fuzzing factors equal to (mf = 2) are employed, and the stopping criterion is set to εf = 0.00001. Data with random outliers and data contaminated by 20% are used to test the model's detection abilities.

Fuzzy clustering methods are utilized to diagnose outliers. The results of the methods are compared based on the standardization property. The best methods have the lowest objective function value among all the values and the least number of iterations required to complete the algorithm, as indicated in Tables 1 and 2.

The methods with the best performance in identifying outliers are shown in Pictures 1 to 48, which also compare the methods on the basis of the clustering of outliers within small or individual groups or whether the distance between these values and the cluster center is twice the average distance in Eq. (14). The objective function and iteration for FCM, PCM 1, and PCM 2 are given in Eqs. (1), (7), and (8), respectively. The objective function and iteration for the proposed method are given as FCM-Stand and PCM-Stand (1) and (2).

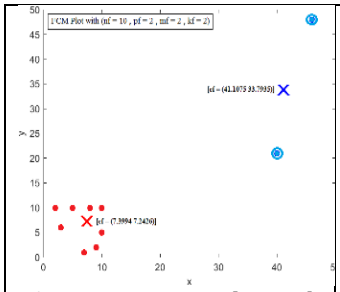
**Table 1.** Results of objective Function (Objf) and Iteration (Itr.) for the FCM, and FCM-Stand

| Objective Function + (Iteration) | | | | | |
|---|---|---|---|---|---|
| **mf** | **pf** | **nf** | **kf** | **FCM (Itr.)** | **FCM-Stand (Itr.)** |
| **2** | **2** | **10** | **2** | 53.85163 (17) | 3.76262 (18) |
| | | | **3** | 21.49420 (72) | 1.89988 (72) |
| | | | **4** | 14.00260 (26) | 1.26552 (25) |
| | | **25** | **2** | 98.32256 (11) | 7.41345 (10) |
| | | | **3** | 73.12346 (81) | 5.56879 (78) |
| | | | **4** | 35.72028 (47) | 3.58532 (71) |
| | | **50** | **2** | 230.05969 (20) | 17.98765 (21) |
| | | | **3** | 169.22769 (100) | 13.29191 (78) |
| | | | **4** | 102.11792 (54) | 8.22219 (52) |
| | | **100** | **2** | 455.85862 (20) | 34.32070 (20) |

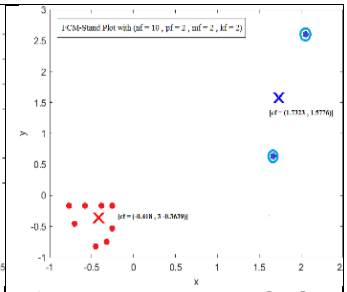
| | | | | | |
|---|---|---|---|---|---|
| | | | **3** | 285.73112 (61) | 21.59256 (59) |
| | | | **4** | 229.07762 (64) | 18.39303 (64) |
| | **3** | **10** | **2** | 49.42499 (15) | 3.28209 (14) |
| | | | **3** | 21.16038 (100) | 1.60507 (35) |
| | | | **4** | 22.83490 (28) | 1.79203 (100) |
| | | **25** | **2** | 122.61011 (15) | 11.73828 (15) |
| | | | **3** | 77.93957 (93) | 5.47478 (100) |
| | | | **4** | 71.11207 (100) | 6.01448 (100) |
| | | **50** | **2** | 279.41437 (13) | 22.17211 (13) |
| | | | **3** | 166.61285 (30) | 14.86833 (31) |
| | | | **4** | 135.04863 (38) | 11.09680 (100) |
| | | **100** | **2** | 577.73959 (18) | 44.69441 (18) |
| | | | **3** | 375.51846 (51) | 29.22223 (51) |
| | | | **4** | 284.37945 (47) | 22.53230 (57) |

**Table 2.** Results of objective Function (Objf) and Iteration (Itr.) with (mf=2) for the PCM (1), PCM (2), and PCM-Stand (1), PCM-Stand (2) methods

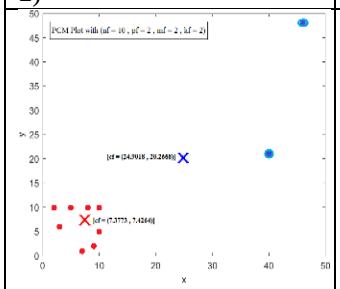
| Objective Function + (Iteration) | | | | | | |
|---|---|---|---|---|---|---|
| **pf** | **nf** | **kf** | **PCM (1) (Itr.)** | **PCM (2) (Itr.)** | **PCM-Stand (1) (Itr.)** | **PCM-Stand (2) (Itr.)** |
| **2** | **10** | **2** | 135.51401 (5) | 17.82280 (5) | 9.60014 (2) | 1.27452 (2) |
| | | **3** | 96.74961 (10) | 7.45376 (10) | 8.55315 (2) | 0.65460 (2) |
| | | **4** | 64.41340 (15) | 3.80574 (15) | 5.81542 (9) | 0.34287 (9) |
| | **25** | **2** | 214.09119 (4) | 16.92276 (4) | 16.08629 (3) | 1.27629 (3) |
| | | **3** | 315.56463 (2) | 12.03916 (2) | 24.05765 (2) | 0.91519 (2) |
| | | **4** | 214.12849 (10) | 4.64788 (10) | 21.85994 (2) | 0.78655 (2) |
| | **50** | **2** | 580.96139 (6) | 66.42676 (6) | 45.58376 (3) | 5.24185 (3) |
| | | **3** | 1046.5602 (100) | 144.2312 (100) | 81.68910 (43) | 11.32936 (43) |
| | | **4** | 904.85297 (13) | 97.21764 (13) | 72.55099 (5) | 7.73194 (5) |
| | **100** | **2** | 1124.8734 (4) | 126.9579 (4) | 84.70946 (3) | 9.56380 (3) |
| | | **3** | 1304.4139( 62) | 49.69549 (62) | 98.57098 (36) | 3.84503 (36) |
| | | **4** | 2030.4466 (27) | 244.3086 (27) | 163.0278 (13) | 19.61653 (13) |
| **3** | **10** | **2** | 114.00961 (3) | 11.50311 (3) | 7.46324 (2) | 0.72152 (2) |
| | | **3** | 81.75323 (27) | 0.14335 (27) | 6.02730 (2) | 0.15812 (2) |
| | | **4** | 158.62871 (5) | 0.94722 (5) | 12.44678 (13) | 0.19936 (13) |
| | **25** | **2** | 264.72867 (5) | 27.46438 (5) | 25.14949 (3) | 2.58688 (3) |
| | | **3** | 324.43290 (2) | 13.43611 (2) | 22.91929 (2) | 0.93369 (2) |
| | | **4** | 682.06599 (100) | 84.12919 (100) | 56.96522 (53) | 7.19543 (53) |
| | **50** | **2** | 679.12378 (5) | 82.10532 (5) | 53.66082 (3) | 6.41382 (3) |
| | | **3** | 757.06811 (25) | 41.57723 (25) | 67.61528 (22) | 3.71479 (22) |
| | | **4** | 1259.8539 (33) | 149.3852 (33) | 107.9391 (100) | 13.00367 (100) |
| | **100** | **2** | 1471.1337 (5) | 187.9107 (5) | 113.8714 (3) | 14.54920 (3) |
| | | **3** | 1743.66014 (2) | 106.6016 (2) | 135.6607 (2) | 8.28563 (2) |
| | | **4** | 2765.5832 (3) | 332.1430 (3) | 218.8192 (2) | 26.31089 (2) |



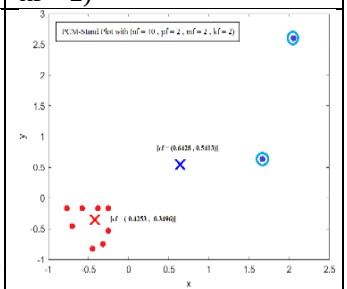
**Picture 1.** FCM Plot with (nf = 10, pf = 2, mf = 2, kf = 2)



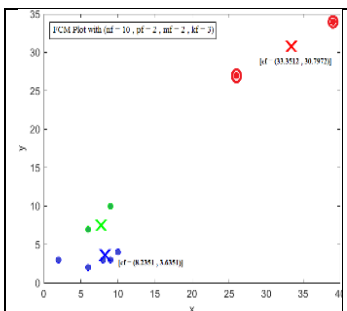
**Picture 2.** FCM-Stand Plot with (nf = 10, pf = 2, mf = 2, kf = 2)



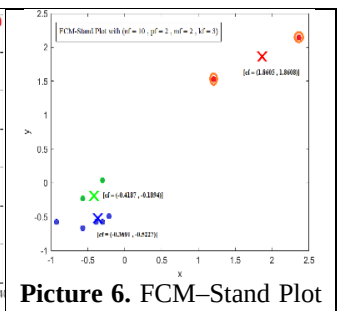
**Picture 3.** PCM Plot with (nf = 10, pf = 2, mf = 2, kf = 2)



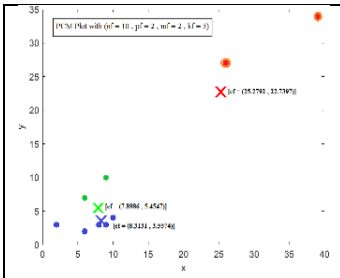
**Picture 4.** PCM-Stand Plot with (nf = 10, pf = 2, mf = 2, kf = 2)



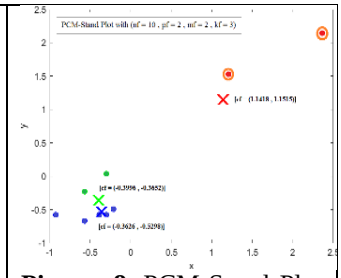
**Picture 5.** FCM Plot with (nf =10, pf = 2, mf = 2, kf = 3)



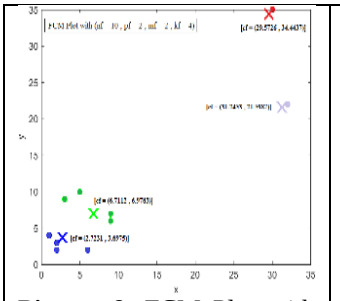
**Picture 6.** FCM–Stand Plot with (nf = 10, pf = 2, mf = 2, kf = 3)



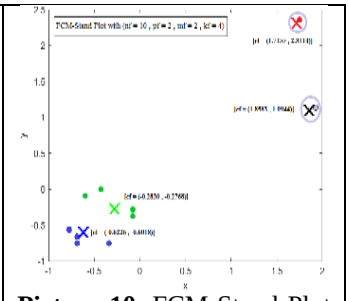
**Picture 7.** PCM Plot with (nf = 10, pf = 2, mf = 2, kf = 3)



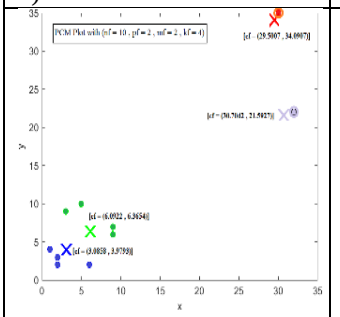
**Picture 8.** PCM-Stand Plot with (nf = 10, pf = 2, mf = 2, kf = 3)



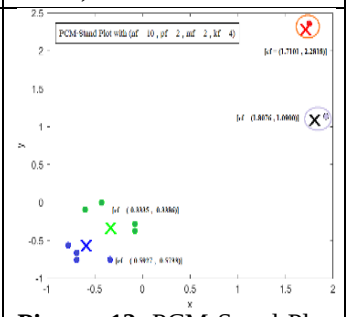
**Picture 9.** FCM Plot with (nf = 10, pf = 2, mf = 2, kf = 4)



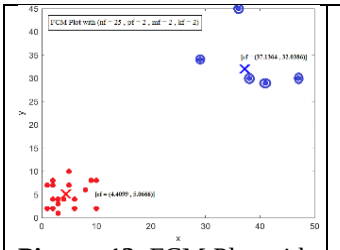
**Picture 10.** FCM-Stand Plot with (nf = 10, pf = 2, mf = 2, kf = 4)



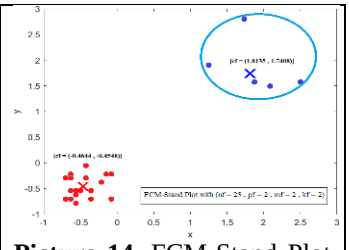
**Picture 11.** PCM Plot with (nf = 10, pf = 2, mf = 2, kf = 4)



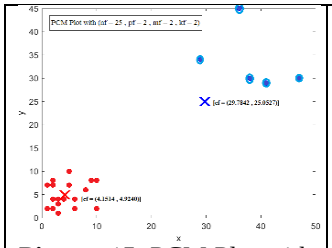
**Picture 12.** PCM-Stand Plot with (nf = 10, pf = 2, mf = 2, kf = 4)



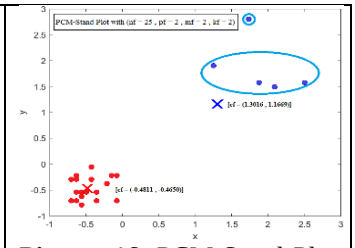
**Picture 13.** FCM Plot with (nf = 25, pf = 2, mf = 2, kf = 2)



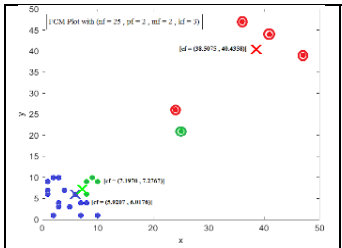
**Picture 14.** FCM-Stand Plot with (nf = 25, pf = 2, mf = 2, kf = 2)



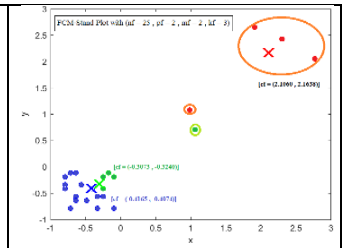
**Picture 15.** PCM Plot with (nf = 25, pf = 2, mf = 2, kf = 2)



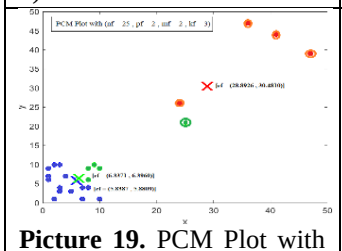
**Picture 16.** PCM-Stand Plot with (nf = 25, pf = 2, mf = 2, kf = 2)



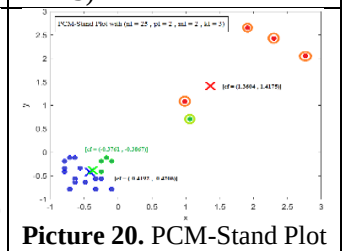
**Picture 17.** FCM Plot with (nf = 25, pf = 2, mf = 2, kf = 3)



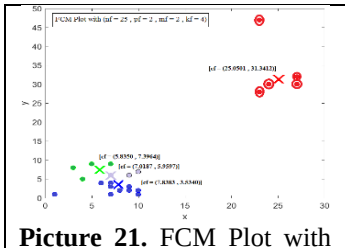
**Picture 18.** FCM-Stand Plot with (nf = 25, pf = 2, mf = 2, kf = 3)



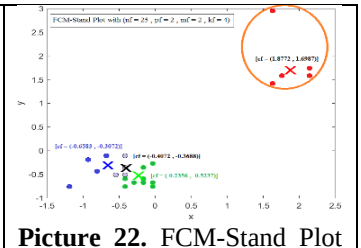
**Picture 19.** PCM Plot with (nf = 25, pf = 2, mf = 2, kf = 3)



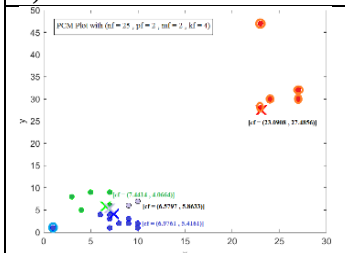
**Picture 20.** PCM-Stand Plot with (nf = 25, pf = 2, mf = 2, kf = 3)



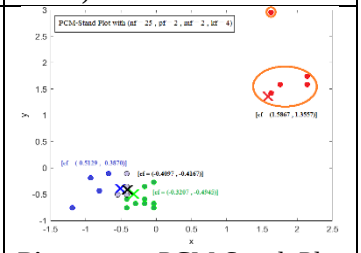
**Picture 21.** FCM Plot with (nf = 25, pf = 2, mf = 2, kf = 4)



**Picture 22.** FCM-Stand Plot with (nf = 25, pf = 2, mf = 2, kf = 4)



**Picture 23.** PCM Plot with (nf = 25, pf = 2, mf = 2, kf = 4)

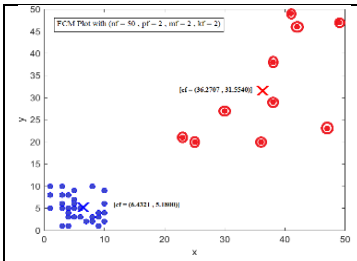


**Picture 24.** PCM-Stand Plot with (nf = 25, pf = 2, mf = 2, kf = 4)

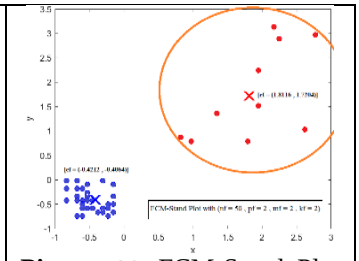
P- ISSN 1991-8941 E-ISSN 2706-6703
2025,(19), (01):260 – 268

*Journal of University of Anbar for Pure Science (JUAPS)*

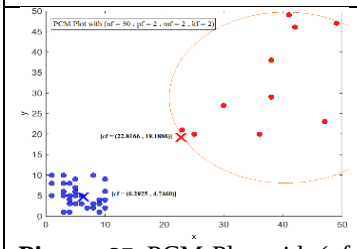
Open Access



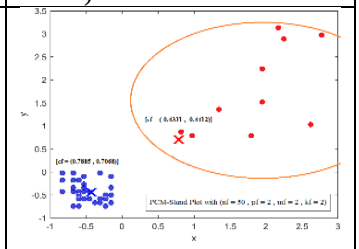
**Picture 25.** FCM Plot with (nf = 50, pf = 2, mf = 2, kf = 2)



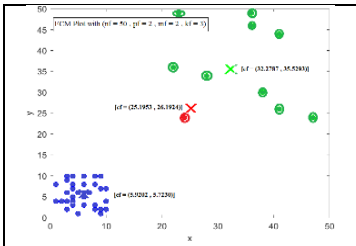
**Picture 26.** FCM-Stand Plot with (nf = 50, pf = 2, mf = 2, kf = 2)



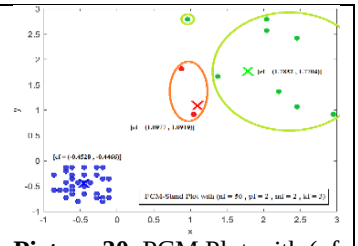
**Picture 27.** PCM Plot with (nf = 50, pf = 2, mf = 2, kf = 2)



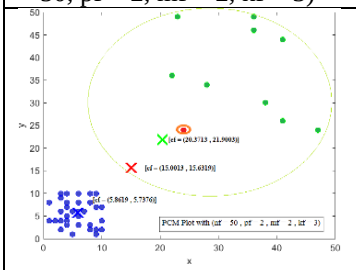
**Picture 28.** PCM-Stand Plot with (nf = 50, pf = 2, mf = 2, kf = 2)



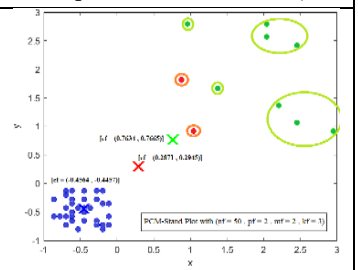
**Picture 29.** PCM Plot with (nf = 50, pf = 2, mf = 2, kf = 3)



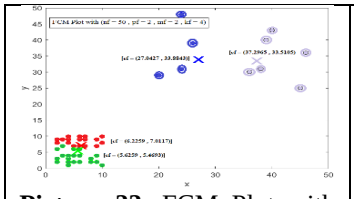
**Picture 30.** PCM Plot with (nf = 50, pf = 2, mf = 2, kf = 3)



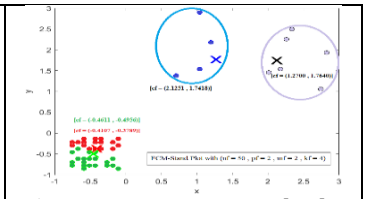
**Picture 31.** PCM Plot with (nf = 50, pf = 2, mf = 2, kf = 3)



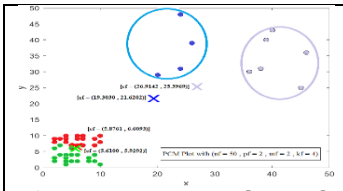
**Picture 32.** PCM-Stand Plot with (nf = 50, pf = 2, m =2, kf=3)



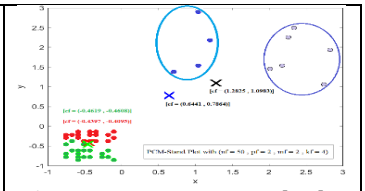
**Picture 33.** FCM Plot with (nf=50, pf = 2, mf = 2, kf = 4)



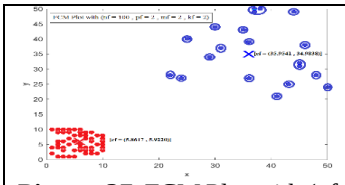
**Picture 34.** FCM-Stand Plot with (nf=50, pf =2, mf =2, kf =4)



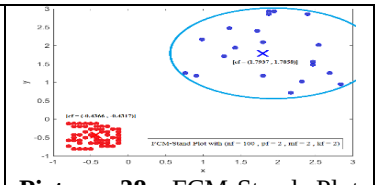
**Picture 35.** PCM Plot with (nf = 50, pf = 2, mf = 2, kf = 4)



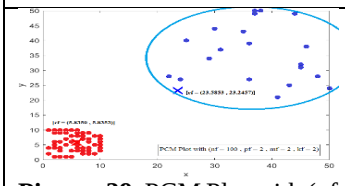
**Picture 36.** PCM-Stand Plot with (nf=50, pf =2, mf =2, kf =4)



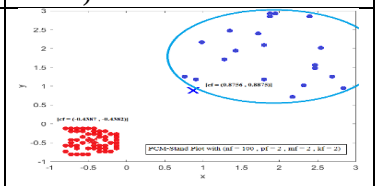
**Picture 37.** FCM Plot with (nf = 100, pf = 2, mf = 2, kf = 2)



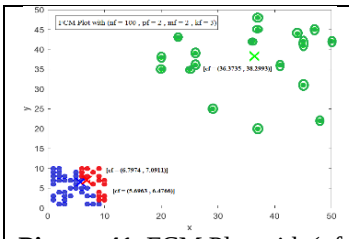
**Picture 38.** FCM-Stand Plot with (nf = 100, pf = 2, mf = 2, kf = 2)



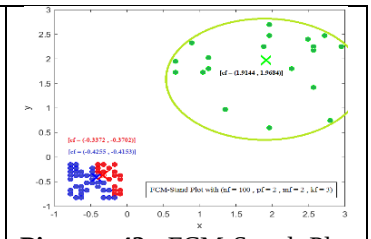
**Picture 39.** PCM Plot with (nf = 100, pf = 2, mf = 2, kf = 2)



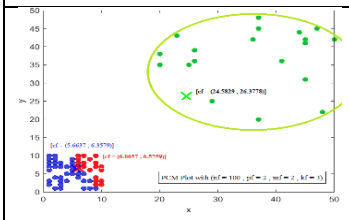
**Picture 40.** PCM-Stand Plot with (nf = 100, pf = 2, mf = 2, kf = 2)



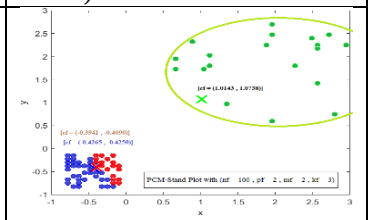
**Picture 41.** FCM Plot with (nf = 100, pf = 2, mf = 2, kf = 3)



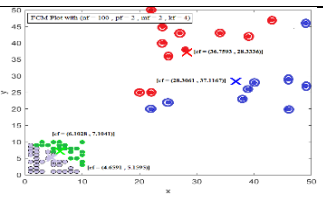
**Picture 42.** FCM-Stand Plot with (nf = 100, pf = 2, mf = 2, kf = 3)



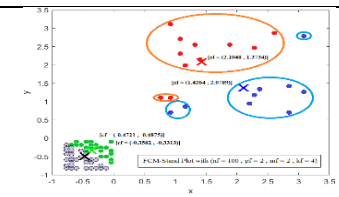
**Picture 43.** PCM Plot with (nf = 100, pf = 2, mf = 2, kf = 3)



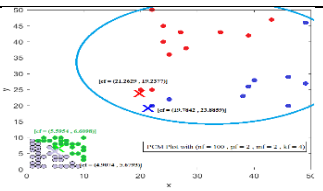
**Picture 44.** PCM-Stand Plot with (nf = 100, pf = 2, mf = 2, kf = 3)



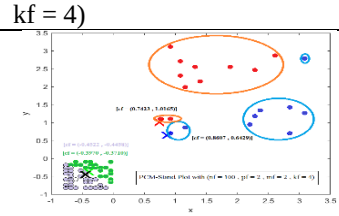
**Picture 45.** FCM Plot with (nf = 100, pf = 2, mf = 2, kf = 4)



**Picture 46.** FCM-Stand Plot with (nf = 100, pf = 2, mf = 2, kf = 4)



**Picture 47.** PCM Plot with (nf = 100, pf = 2, mf = 2, kf = 4)



**Picture 48.** PCM-Stand Plot with (nf = 100, pf = 2, mf = 2, kf = 4)

## Conclusions

The performance of the proposed technique is compared with that of other approaches. The result implies that when fuzzy clustering methods FCM and PCM are applied, standardization is an essential initial phase to increasing the precision of outlier diagnosis. Substantial performance improvement and consistency in diagnostic outcomes can be attained by the implementation of standardization. Most of the time, uniformity is the ideal option because it can convert data into a consistent range, thus improving the precision of the clustering criterion.

1- In the absence of standardization, FCM performs moderately well in identifying outliers. The number of iterations increases in both cases as the sample sizes and cluster numbers increase. Meanwhile, the values of objective functions begin to decrease as the number of clusters increases (pf = 2, 3).
2- When the sample sizes and cluster numbers increase, the FCM method with the use of the standardization approach (FCM-Stand) exerts a positive effect on decreasing the degrees of objective functions and the number of iterations when (pf = 2, 3) compared with the FCM method.
3- When the two different objective function formulas are used, the PCM method performs differently. In the first case, it performs well when the objective function in Eq. (8) that represents the method PCM 2 is used. It does not perform as well when the objective function in Eq. (7) that represents the method PCM 1 is employed.
4- Excellent results are obtained together with high diagnostic accuracy and a noticeable improvement in identifying outliers when the standardization approach with the PCM method is adopted. Notable decrements in the values of the objective functions and the number of iterations are achieved when the number of clusters increases and when pf = 2, 3.
5- The PCM 2 method performs better than the FCM method in the cases with and without using the standardization strategy.
6- Dealing with standardized data makes the distance criteria used in FCM and PCM increasingly precise. For the best results, the model's performance should be regularly assessed, and the standardization procedures should be improved when necessary.

## Acknowledgments

We appreciate the support provided by the Department of Mathematics, College of Science, Mustansiriyah University. We are also grateful to your esteemed journal for accepting this study for publication.

## Conflict of Interest

There are no conflicts of interest.

## References

[1] zadeh, L. A. (1965). fuzzy set, information and control. 338-353.
[2] zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning, Part 1. Information Science, 199-249.
[3] Rhee, F. C., & Hwang, C. (2001). type-2 fuzzy c-means clustering algorithm. in EEE 9th IFSA World Congress and 20th NAFIPS Conf.
[4] Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms, Plenum.
[5] Dunn, J. (1974). A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3, 32-57.
[6] Gustafson, D. E., & Kessel, W. C. (1979). Fuzzy clustering with a fuzzy covariance matrix.. in In Proc. of IEEE Conference on Decision and Control

including the 17th Symposium on Adaptive Processes, San Diego.

[7] Ōhashi, Y. (1984). Fuzzy Clustering and Robust Estimation. Paper in University Tokyo Hospital.

[8] Dave, R. N. (1991). Characterization and Detection of Noise in Clustering. *Pattern Recognition Letters*, 12, 657-664.

[9] Dave, R., & Krishnapuram N. (1997). Robust Clustering Methods: A Unified View. *IEEE Trans. Fuzzy Systems*, 5(2), 270-293.

[10] Bandemer, H., & Gottwald, S. (1993). Einfiihrung in Fuzzy-Methoden. (4. Aufl.), Berlin: Akademie Verlag.

[11] Krishnapuram, R., & Keller J. (1993). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2), 98-110.

[12] Krishnapuram, R., & Keller J. (1996). The possibilistic c-means algorithm: insights and recommendations. *IEEE Transactions on Fuzzy Systems*, 4(3), 385-393.

[13] Pal, N. R., Pal, K., & Bezdek, J. C. (1997). A mixed c-means clustering model in *In Proc. of the 6th IEEE Int. Conf. on Fuzzy Systems* (Vol. 3).

[14] Pal, N. R., Pal, K., Keller, J. M., & Bezdek, J. C. (2005). A Possibilistic Fuzzy C-Means Clustering Algorithm," *IEEE Transactions on Fuzzy Systems*, 13(4), 517-530.

[15] Yang, M. S., & Wu, K. L. (2006). Unsupervised possibilistic clustering. *Pattern Recognition*, 39(1), 5-21.

[16] Wu, X., Wu, B., Sun, J., & Fu, H. (2010). Unsupervised possibilistic fuzzy clustering. *J. of Information and Computational Sci.*. 7(5), 1075-1080.

[17] Zgurovsky, M. Z., & Zaychenko, Y. P. (2020). Big Data: Conceptual Analysis and Applications, Springer Nature Switzerland AG.

[18] Höppner, F., Klawonn, F., & Kruse R. (2000). Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition, Wiley IBM PC Series, 1st Edition.

[19] Bezdek, J. C. (1974). Cluster validity with fuzzy sets. *J. Cybernetics*, 3, 58-73.

[20] Filippone, M., Masulli, F., & Rovetta, S. (2007). Possibilistic clustering in feature space. In Int. Workshop on Fuzzy Logic and Applications. *Springer, Berlin, Heidelberg*, 219-226.

[21] Timm, H., Borgelt, C., Döring, C., & Kruse R. (2004). An extension to possibilistic fuzzy cluster analysis. *Fuzzy Sets and Systems*, 147(1), 3-16.

[22] Duda, R., & Hart, P. (1973). Pattern Classification and Scene Analysis, NEWYORK: WILEY.

# تقنية جديدة لتشخيص القيم الشاذة باستخدام العنقدة الضبابية

**سجى محمد سكران ، وفاء سيد حسنين***

قسم علوم الرياضيات، كلية العلوم، جامعة المستنصرية، بغداد، العراق

**الخلاصة:**

القيم الشاذة داخل مجموعة البيانات عبارة عن قيم تختلف بشكل كبير عن بقية البيانات. ويمكن أن تُعزى نقاط البيانات غير النمطية هذه الى مجموعة من العوامل، مثل اخطاء القياس، ومشاكل ادخال البيانات، والاختلافات الطبيعية في البيانات. وتعد عملية اكتشاف وتعديل القيم الشاذة أمرًا ضروريًا لضمان سلامة التحليلات الإحصائية وتجنب الحصول على نتائج مضللة. يمكن ملاحظة هذه القيم المتطرفة إما عند نقاط عالية جدًا أو منخفضة جدًا، ولها تأثير ملحوظ على المقاييس الإحصائية مثل المتوسط والتباين. وتركز العديد من تقنيات التشخيص على نقاط التمركز المؤثرة والمسافات بين المجموعات للكشف عن هذه القيم غير الطبيعية. وبصورة خاصة تم في هذه الدراسة استخدام تقنيات المجموعة الضبابية لتحديد القيم المتطرفة داخل مجموعة البيانات من حيث اقتراح تقنية بديلة للكشف عن القيم المتطرفة باستخدام مفهوم ونهج التقييس باستخدام طرائق المجموعة الضبابية في الكشف عن القيم الشاذة في مجموعة البيانات. وكذلك تم مقارنة أداء الطريقة المقترحة بالطرق الأخرى باستخدام المحاكاة.

الكلمات المفتاحية :متوسطات –c الضبابية، متوسطات –c الاحتمالية، التقييس، دالة الهدف، المحاكاة.