Partition Coefficient and Partition Entropy in Fuzzy C Means Clustering

Rohit Kumar Verma [a], Rakesh Tiwari [b] and Pratik Singh Thakur [b*]

[a] *Department of Mathematics, Govt. Chandulal Chandrakar Arts and Science College, Patan, C.G., 491111, India.*
[b] *Department of Mathematics, Govt. VYT PG. Autonomous College, Durg, C.G., 491001, India.*

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Original Research Article

ABSTRACT

This paper offers a comprehensive exploration of partition validation functions, specifically focusing on partition coefficient and partition entropy within the realm of fuzzy clustering—an influential approach in the field of clustering datasets. While fuzzy clustering facilitates the classification of data points into multiple clusters, the pivotal tasks of determining the optimal number of clusters and evaluating the validity of the resultant clusters pose inherent challenges. The study addresses these challenges, contributing to the broader understanding of effective fuzzy clustering methodologies.

*Corresponding author: E-mail: spratik343@gmail.com

1 INTRODUCTION

Fuzzy c-means (FCM) clustering is a widely used technique for partitioning data into clusters based on similarity measures. The technique extends the traditional k-means algorithm by allowing data points to belong to more than one cluster, with a degree of membership represented by a fuzzy partition matrix [1]. FCM clustering has been applied in various domains, including image segmentation, data mining, and bioinformatics, among others. Several studies have investigated the effectiveness of FCM clustering and its variants, such as the adaptive fuzzy c-means (AFCM) algorithm and the fuzzy possibilistic c-means (FPCM) algorithm, in different applications (see [2, 3, 4, 5]).

For example, in image segmentation, FCM clustering has been used to segment brain MRI images [6] and satellite images [7]. In bioinformatics, FCM clustering has been applied to gene expression data analysis [8] and protein structure prediction [9]. In addition, several studies have proposed modifications to the FCM algorithm to improve its performance, such as the use of kernel-based fuzzy c-means clustering [10] and the incorporation of partition coefficient and partition entropy measures [11]. Overall, FCM clustering and its variants continue to be an active area of research, with ongoing efforts to improve its accuracy, efficiency, and applicability.

Clustering validity functions are important tools for evaluating the quality of clustering results and selecting the appropriate number of clusters. These functions provide quantitative measures of the clustering performance, based on the distribution of the data points and the distance between clusters. Common clustering validity functions include the silhouette coefficient, the Calinski-Harabasz index, and the Davies-Bouldin index. These functions have been applied in various domains, including image analysis, bioinformatics, and social network analysis, among others. Several studies have investigated the use of clustering validity functions for improving clustering performance, such as the work of [12, 13, 14, 15, 16]. Overall, clustering validity functions are essential for evaluating and improving clustering algorithms, and their use can lead to more accurate and reliable clustering results.

This study looks at two things, partition coefficient (PC) and partition entropy (PE), to understand how well our clusters are separated and if they're clear or fuzzy. If the PC is small, it might mean the data is tricky, but if it's big, our clusters are well-defined. Similarly, a small PE suggests clear clusters, while a big PE could mean trouble. This journey through fuzzy c-means clustering explores these ideas, showing how each measurement tells us something important about cluster quality and makes us rethink what we thought we knew about getting the best clusters.

2 THE FUZZY C-MEANS ALGORITHM

Fuzzy C-Means (FCM), sometimes known as fuzzy K-Means, is a fuzzy variant of the K-Means algorithm that was proposed by Bezdek [17, 18]. The least-square error criteria is the foundation of FCM. FCM beats K-Means because it assigns each pattern to each cluster with a certain level of membership (i.e. fuzzy clustering). This works better in practical settings where there are some cluster overlaps in the data set. The FCM optimises the following objective function:

$$J_{FCM} = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{j,i}^{q} d(x_i, m_j) \qquad (2.1)$$

where q denotes the fuzziness exponent, and $q \geq 1$. The algorithm becomes more fuzzy as the value of q increases; $u_{j,i}$ is the membership value for the ith pattern in the jth cluster satisfying the following constraints:

1. $u_{j,i} \geq 0,\ i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$

2. $\sum_{j=1}^{k} u_{j,i} = 1,\ i = 1, 2, \ldots, n$

For FCM the membership function is denifend as

$$u(m_j|x_i) = \frac{\|x_i - m_j\|^{-2/(q-1)}}{\sum_{j=1}^{k} \|x_i - m_j\|^{-2/(q-1)}} \qquad (2.2)$$

and weight function is defined as

$$w(x_i) = 1 \qquad (2.3)$$

As a result, FCM features a constant weight function as well as a soft membership function. FCM generally outperforms K-Means [19] and is less impacted by the existence of data uncertainty [20]. The user must yet define the number of clusters in the data set, just like in K-Means (see [21]. Additionally, it could reach local optimum [22, 14, 23].

3 CLUSTERING VALIDITY INDEX

A partition index for hard clustering was proposed by Dunn in 1974 [24]. The first Fuzzy Clustering Validity Index (FCVI) was then created using this method, and Bezdek [25] offered the partition coefficient [26] (V_{PC}) and partition entropy [27] (V_{PE}) as fuzzy clustering validity functions, as given in Eq. 3.1–3.2. V_{PC} and V_{PE} are both valid for both the maximum and minimum values. These are only membership based validity functions, where $V_{PC} \in [0, 1]$ and $V_{PE} \in [0, log_a c]$. In order to more effectively describe the fuzziness of data samples, V_{PC} introduces the fuzzy weighted m. V_{PC} and V_{PE} have a straightforward structure and need less work, but they will alter monotonically when more clusters are added. As a result, V_{PC} and V_{PE} have a limited capacity to handle data sets with intricate architecture [27] - [36].

Definition 3.1 (Partition Coefficient). [26].

$$V_{PC} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} u_{j,i}^{m} \qquad (3.1)$$

Definition 3.2 (Partition Entropy). [27]

$$V_{PE} = -\frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} u_{j,i} \log_a(u_{j,i}) \qquad (3.2)$$

4 IMPLIMENTATION OF METHODOLOGY

A random data set generated using rand(250,2) in python. then sci-kit fuzzy library is used to perform fcm on that data set. Then PC and PE are calculated. We get results for 2,3 and 4 number of clusters.

As seen in Fig. 1, PC is high and PE is low due to the sparse overlap of clusters. Clusters are slightly overlapping in Fig. 2, and PE is high. Clusters are heavily overlapped in Fig. 3, PC is low and PE is high.

Table 1. Value of PC and PE for random Data Set

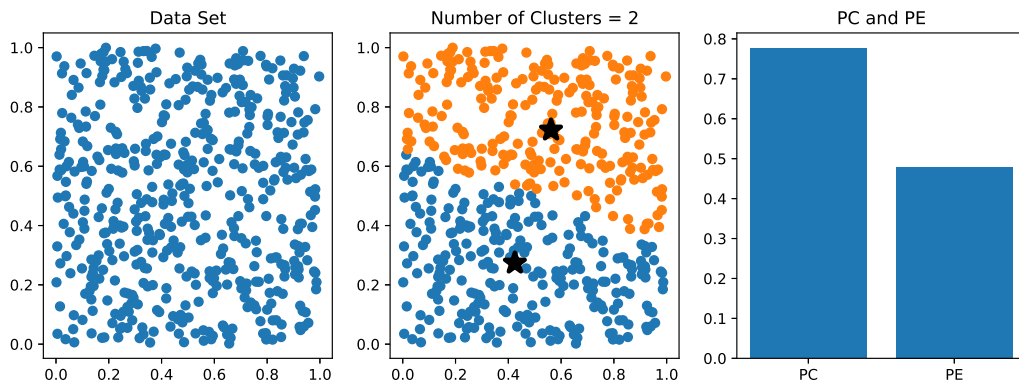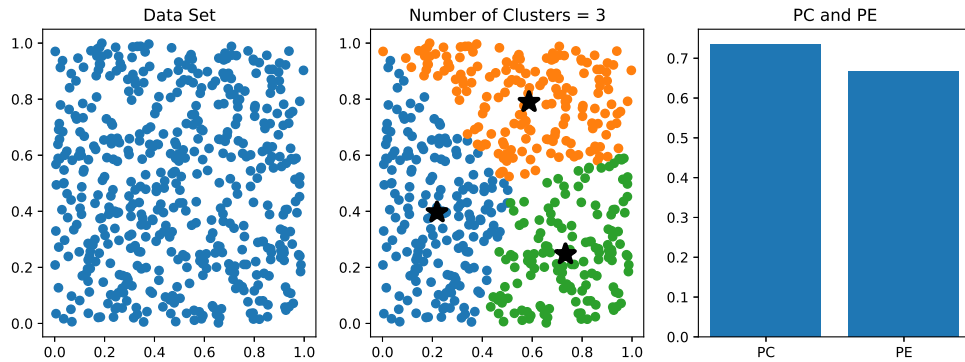| Number of clusters (k)	Partition Coefficient	Partition Entropy
2	0.7760268733421876	0.47872665124718355
3	0.7361233295978787	0.6668485796708312
4	0.7214981079291537	0.7850044547581662

Fig. 1. fcm for k = 2
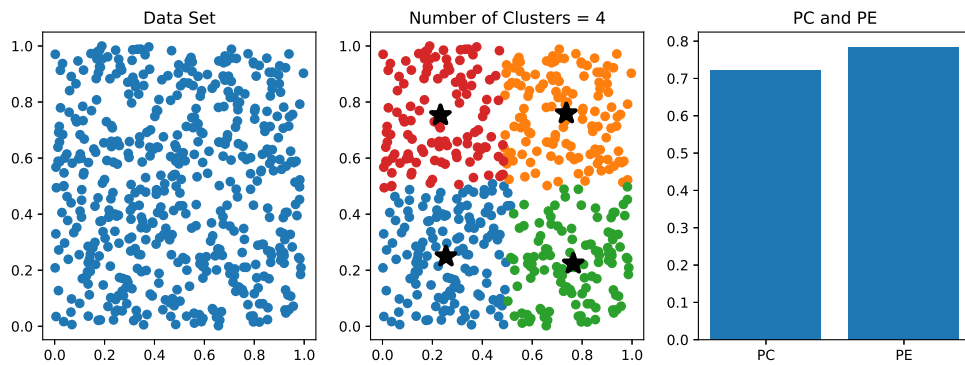
Fig. 2. fcm for k = 3

Fig. 3. fcm for k = 4

5 CONCLUSION

Partition coefficient (PC) and partition entropy (PE) are measures used to evaluate the quality of clustering results in fuzzy c-means clustering. When the value of PC is larger, it indicates that the degree of overlap between the clusters is smaller and the clusters are better separated. On the other hand, when the value of PE is larger, it indicates that the degree of fuzziness of the clusters is higher and the clusters are more overlapping.

For smaller values of PC, the clusters tend to be more overlapping, which may indicate that the data is inherently difficult to cluster. However, a smaller value of PC may also indicate that the number of clusters is not appropriate for the data. In contrast, for larger values of PC, the clusters tend to be more well-separated, which can be desirable for some clustering applications.

Similarly, for smaller values of PE, the clusters tend to be less fuzzy and more well-defined. This can be desirable for clustering applications where the data is not inherently difficult to cluster. However, for larger values of PE, the clusters tend to be more fuzzy and overlapping. This may indicate that the number of clusters is not appropriate for the data or that the data is inherently difficult to cluster. Overall, both PC and PE provide valuable insights into the quality of the clustering results in fuzzy c-means clustering, and the choice of which measure to use depends on the specific problem and the desired clustering objectives.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

[1] Zadeh LA. Fuzzy sets. Information and Control. 1965;8(3):338–353.
DOI: https://doi.org/10.1016/S0019-9958(65)90241-X

[2] Ajin V, Kumar LD. Big data and clustering algorithms. 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS). 2016;1–5.

[3] Nayak J, Naik B, Behera Prof. Dr. H. Fuzzy C-means (FCM) clustering algorithm: A decade review from 2000 to 2014. In Smart Innovation, Systems and Technologies. Springer. 2015;32:133–149.
DOI: https://doi.org/10.1007/978-81-322-2208-8_14

[4] Pérez-Ortega J, Almanza-Ortega NN, Romero D. Balancing effort and benefit of K-means clustering algorithms in big data realms. PLoS One. 2018;13(9):e0201874.

[5] Shirkhorshidi AS, Aghabozorgi SR, Wah TY, Herawan T. Big data clustering: A review. Communication Systems and Applications; 2014.

[6] Hua L, Gu Y, Gu X, Xue J, Ni T. A novel brain MRI image segmentation method using an improved multi-view fuzzy c-means clustering algorithm. Frontiers in Neuroscience. 2021;15:662674.

[7] Salah M. Extraction of road centrelines and edge lines from high-resolution satellite imagery using density-oriented fuzzy C-means and mathematical morphology. Journal of the Indian Society of Remote Sensing. 2022;50(7): 1243–1255.

[8] Maraziotis IA. A semi-supervised fuzzy clustering algorithm applied to gene expression data. Pattern Recognition. 2012;45(1):637–648.

[9] Kaur S, Sharma A, Singh P. Performance analysis of fuzzy C-means (FCM) and possibilistic C-means algorithm (PCM) for protein sequence clustering. Proceedings of the International Conference on Innovative Computing& Communication (ICICC); 2021.

[10] Hu G, Du Z. Adaptive kernel-based fuzzy c-means clustering with spatial constraints for image segmentation. International Journal of Pattern Recognition and Artificial Intelligence. 2019;33(01):1954003.

[11] Wang H, Wang J, Wang G. Combination evaluation method of fuzzy c-mean clustering validity based on hybrid weighted strategy. IEEE Access: Practical Innovations, Open Solutions. 2021;9:27239–27261.

[12] Bellman R, Kalaba R, Zadeh L. Abstraction and pattern classification. Journal of Mathematical Analysis and Applications. 1966;13(1):1–7.
DOI: https://doi.org/10.1016/0022-247X(66)90071-0

[13] Charrad M, Ghazzali N, Boiteau V, Niknafs A. Determining the number of clusters using NbClust package. MSDM. 2014;1.

[14] Modak S. Distinction of groups of gamma-ray bursts in the BATSE catalog through fuzzy clustering. Astronomy and Computing. 2021;34: 100441.

[15] Ruspini EH, Bezdek JC, Keller JM. Fuzzy clustering: A historical perspective. IEEE Computational Intelligence Magazine. 2019b;14(1):45–55.
DOI: https://doi.org/10.1109/MCI.2018.2881643

[16] Zhang B. Generalized K-harmonic means boosting in unsupervised learning. Hewlett Packard Laboratory. 2000;1:1–12.

[17] Bezdek JC. A convergence theorem for the fuzzy ISODATA clustering algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-2(1). 1980;1–8.
DOI: https://doi.org/10.1109/TPAMI.1980.4766964

[18] Bezdek JC. Objective function clustering. In Pattern recognition with fuzzy objective function algorithms. Springer US. 1981;43–93.
DOI: https://doi.org/10.1007/978-1-4757-0450-1_3

[19] Hamerly GJ. Learning structure and concepts in data through data clustering. University of California, San Diego; 2003.

[20] Liew A, Leung S, Lau W. Fuzzy image clustering incorporating spatial continuity. IEE Proceedings-Vision, Image and Signal Processing. 2000;147(2):185–192.

[21] Naldi MC, Campello RJ. Comparison of distributed evolutionary k-means clustering algorithms. Neurocomputing. 2015;163:78–93.

[22] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. Acm Computing Surveys. 1999;31(3):264–323.
DOI: https://doi.org/10.1145/331499.331504

[23] Yang MS. A survey of fuzzy clustering. Mathematical and Computer Modelling. 1993;18(11):1–16.
DOI: https://doi.org/10.1016/0895-7177(93)90202-A

[24] Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics. 1973;3(3):32–57.
DOI: https://doi.org/10.1080/01969727308546046

[25] Bezdek J, Pal N. Some new indexes of cluster validity. IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics : A Publication of the IEEE Systems, Man, and Cybernetics Society. 1998;28:301–315.
DOI: https://doi.org/10.1109/3477.678624

[26] Bezdek JC. Numerical taxonomy with fuzzy sets. Journal of Mathematical Biology. 1974;1(1):57–71.

[27] James C. Bezdekâ€. Cluster validity with fuzzy sets. Journal of Cybernetics. 1973;3(3):58–73.
DOI: https://doi.org/10.1080/01969727308546047

[28] Ever-Hadani S. Applications of cluster analysis algorithm to geostatistical series. Regional Science and Urban Economics. 1980;10(1):123–151.
DOI: https://doi.org/10.1016/0166-0462(80)90052-6

[29] Ghosh S, Dubey SK. Comparative analysis of K-means and fuzzy C-means algorithms. International Journal of Advanced Computer Science and Applications. 2013;4(4).
DOI: https://doi.org/10.14569/IJACSA.2013.040406

[30] Hamerly G, Elkan C. Alternatives to the K-means algorithm that find better clusterings. Proceedings of the Eleventh International Conference on Information and Knowledge Management. 2002;600–607.
DOI: https://doi.org/10.1145/584792.584890

[31] Lee GM, Gao X. A hybrid approach combining fuzzy c-means-based genetic algorithm and machine learning for predicting job cycle times for semiconductor manufacturing. Applied Sciences. 2021;11(16):7428.
DOI: https://doi.org/10.3390/app11167428

[32] Lee SJ, Song DH, Kim KB, Park HJ. Efficient fuzzy image stretching for automatic ganglion cyst extraction using fuzzy C-means quantization. Applied Sciences. 2021;11(24):12094.
DOI: https://doi.org/10.3390/app112412094

[33] MacQueen J. Classification and analysis of multivariate observations. 5th Berkeley Symp. Math. Statist. Probability. 1967;281–297.

[34] Omran M, Engelbrecht A, Salman A. An overview of clustering methods. Intelligent Data Analysis. 2007;11:583–605.
DOI: https://doi.org/10.3233/IDA-2007-11602

[35] Ruspini EH. A new approach to clustering. Information and Control. 1969;15(1):22–32.
DOI: https://doi.org/10.1016/S0019-9958(69)90591-9

[36] Ruspini EH, Bezdek JC, Keller JM. Fuzzy clustering: A historical perspective. IEEE Computational Intelligence Magazine. 2019a;14(1):45–55.