Original Research Article

Partition Coefficient and Partition Entropy in Fuzzy C Means Clustering

Abstract

This paper provides an overview of partition validation functions such as partition coefficient and partition entropy used in fuzzy clustering, a popular method for clustering data sets. Fuzzy clustering enables classification of data points into multiple clusters, but selecting the appropriate number of clusters and assessing the validity of the resulting clusters can be challenging.

MSC *: 62H30; 90C70; 68W40; 91C20.*
Keywords *: fuzzy clustering, validation functions, clustering validation, partition coefficient, partition entropy.*

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. 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.

For example, in image segmentation, FCM clustering has been used to segment brain MRI images [10] and satellite images [17]. In bioinformatics, FCM clustering has been applied to gene expression data analysis [14] and protein structure prediction [12]. 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 [9] and the incorporation of partition coefficient and partition entropy measures [18]. 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 [16], [6], and [15]. Overall, clustering validity functions are essential for evaluating and improving clustering algorithms, and their use can lead to more accurate and reliable clustering results.

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 [3, 4]. 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:

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

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 i^{th} pattern in the j^{th} 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

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

and weight function is defined as

$$w(x_i) = 1 \tag{2.3}$$

As a result, FCM features a constant weight function as well as a soft membership function. FCM generally outperforms K-Means [8] and is less impacted by the existence of data uncertainty [13]. The user must yet define the number of clusters in the data set, just like in K-Means. Additionally, it could reach local optimum [11].

3 Clustering Validity Index

A partition index for hard clustering was proposed by Dunn in 1974 [7]. The first Fuzzy Clustering Validity Index (FCVI) was then created using this method, and Bezdek [1] offered the partition coefficient [2] (V_{PC}) and partition entropy [5] (V_{PE}) as fuzzy clustering validity functions, as given in Eqs. 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.

Definition 3.1 (Partition Coefficient)**.**

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

Definition 3.2 (Partition Entropy)**.**

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

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,

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

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

and PE is high. Clusters are heavily overlapped in Fig. 3, PC is low and PE is high.

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.

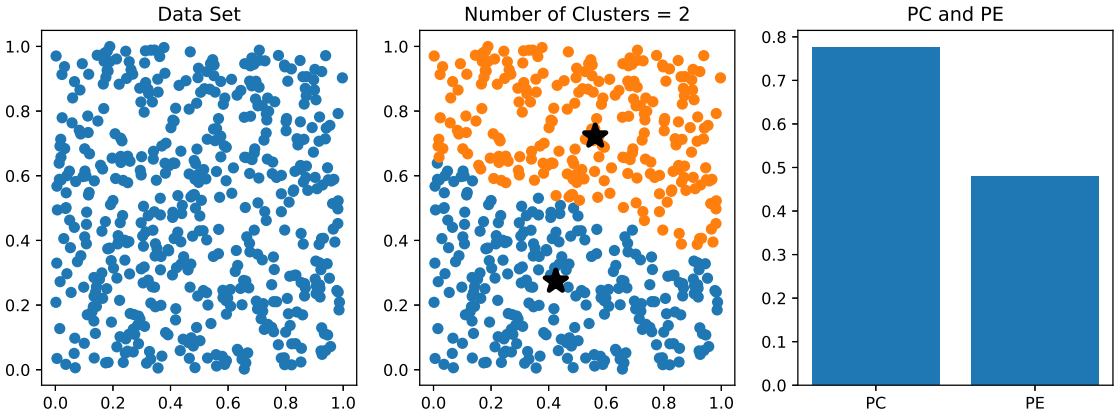

Figure 1: fcm for k = 2

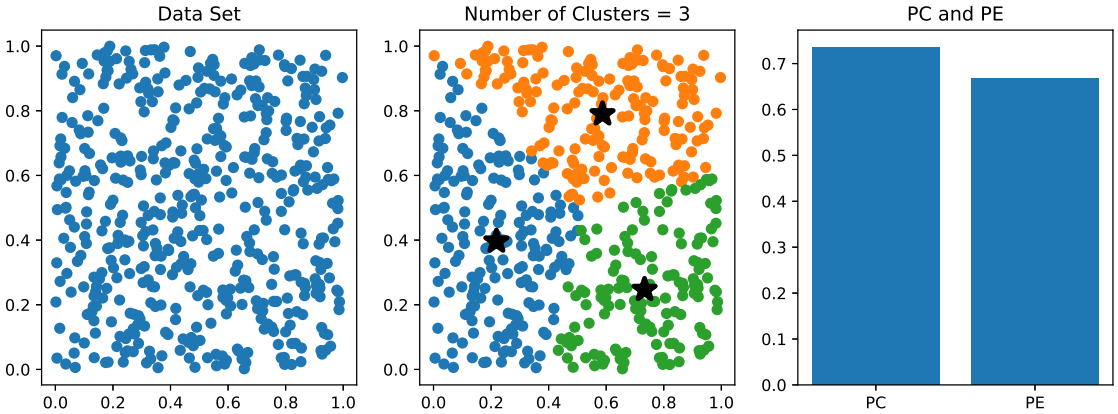

Figure 2: fcm for k = 3

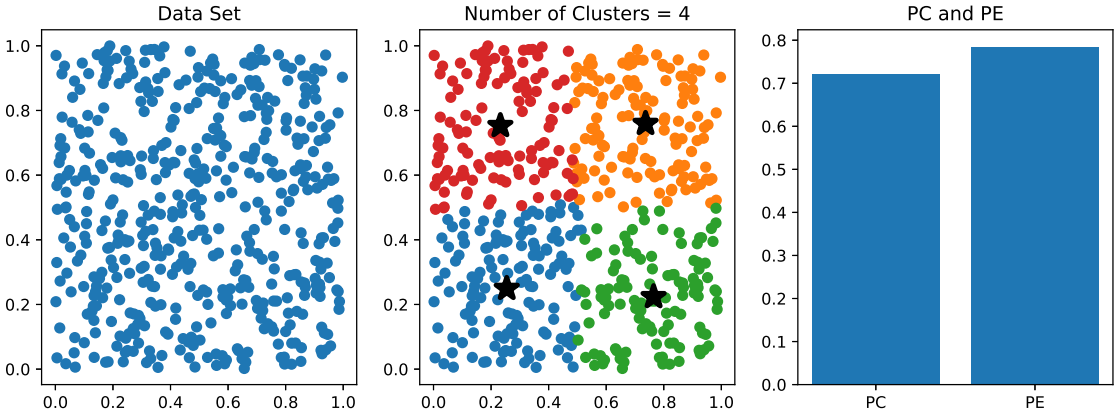

Figure 3: fcm for k = 4

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.

References

[1] Bezdek, J. and Pal, N. (1998). 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*, 28:301–15.

[2] Bezdek, J. C. (1974). Numerical taxonomy with fuzzy sets. *Journal of mathematical biology*, 1(1):57–71.

[3] Bezdek, J. C. (1980). A convergence theorem for the fuzzy isodata clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(1):1–8.

[4] Bezdek, J. C. (1981). *Objective Function Clustering*, pages 43–93. Springer US, Boston, MA.

[5] Bezdekâ€, J. C. (1973). Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3(3):58–73.

[6] Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). Determining the number of clusters using nbclust package. *MSDM*, 2014:1.

[7] Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57.

[8] Hamerly, G. J. (2003). *Learning structure and concepts in data through data clustering*. University of California, San Diego.

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

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

[11] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323.

[12] Kaur, S., Sharma, A., and Singh, P. (2021). Performance analysis of fuzzy c-means (fcm) and possibilistic c-means algorithm (pcm) for protein sequence clustering. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.

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

[14] Maraziotis, I. A. (2012). A semi-supervised fuzzy clustering algorithm applied to gene expression data. *Pattern Recognition*, 45(1):637–648.

[15] Modak, S. (2021). Distinction of groups of gamma-ray bursts in the batse catalog through fuzzy clustering. *Astronomy and Computing*, 34:100441.

[16] Ruspini, E. H., Bezdek, J. C., and Keller, J. M. (2019). Fuzzy clustering: A historical perspective. *IEEE Computational Intelligence Magazine*, 14(1):45–55.

[17] Salah, M. (2022). 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*, 50(7):1243–1255.

[18] Wang, H., Wang, J., and Wang, G. (2021). Combination evaluation method of fuzzy c-mean clustering validity based on hybrid weighted strategy. *IEEE Access*, 9:27239–27261.