Available online at http://scik.org
Eng. Math. Lett. 2024, 2024:1
https://doi.org/10.28919/eml/8403
ISSN: 2049-9337

A COMPARISON BETWEEN THE FUZZY C-MEANS CLUSTERING ALGORITHM AND THE K-MEAN CLUSTERING ALGORITHM

PRATIK SINGH THAKUR[1,*], ROHIT KUMAR VERMA[2], RAKESH TIWARI[1]

[1]Department of Mathematics, Govt. VYT PG Autonomous College, Durg, 491001, India

[2]Department of Mathematics, Govt. Chandulal Chandrakar Arts and Science College, Patan, 491111, India

Abstract. This paper provides brief information about the K-means clustering algorithm and the Fuzzy C-means clustering algorithm. We have defined mathematical expressions for K-means and FCM and have applied the algorithms to a dummy dataset, comparing their differences for various numbers of clusters.

Keywords: clustering; K-means; fuzzy C-means.

2020 AMS Subject Classification: 62H30, 90C70, 68W40, 91C20.

1. INTRODUCTION

In several fields, including bioinformatics, pattern recognition, commerce, and image processing, the clustering of objects according to their qualities has been intensively researched [16, 23]. Additionally, it has become more critical due to the exponential growth of data in several fields of knowledge. A complete dataset is a foundation for their analysis, justification, and decision-making [1, 21].

The goal of clustering is to divide a set of n objects into smaller groups or clusters so that each cluster contains objects with characteristics similar to and distinct from the objects in every

*Corresponding author

E-mail address: spratik343@gmail.com

Received December 16, 2023

other cluster. The description mentioned above fits the definition of a hard or traditional clustering type, frequently associated with the K-Means algorithm [14]. Each item must, however, belong to two or more groups in numerous domains, each with varying degrees of membership [11, 12, 20]. Fuzzy C-Means is the algorithm that makes those mentioned above possible. However, the complexity of FCM exceeds that of K-Means.

Zadeh's [24] fuzzy set theory provides an idea of the membership uncertainty that is described by a membership function. Bellman, Kalaba, and Zadeh [2] proposed the cluster analysis theory, while Ruspini [19] introduced the idea of fuzzy partitioning—more precisely, the fuzzy clustering technique. These papers serve as the foundation for fuzzy clustering research. Dunn [9] expanded the definition of hard grouping in 1973 to include early notions of fuzzy means. Finally, Bezdek [4] expanded on Dunn's strategy to create an unlimited family of Fuzzy C Means algorithms in 1981.

2. BACKGROUND

The terminologies used in the paper are defined in this section, which also gives the reader the background information they need to understand the debate that follows. The following terms are used in this paper:

(1) A pattern (or feature vector), z, is a single object or data point used by the clustering algorithm [10].
(2) A feature (or attribute) is an individual component of a pattern [10].
(3) A cluster is a set of similar patterns, and patterns from different clusters are not similar [5].
(4) Hard (or Crisp) clustering algorithms assign each pattern to one and only one cluster.
(5) Fuzzy clustering algorithms assign each pattern to each cluster with some degree of membership.
(6) A distance measure is a metric used to evaluate the similarity of patterns [10].

Definition 2.1 (Clustering Problem)**.** [17] The following is a formal definition of the clustering problem.

Let $X = \{x_1, x_2, \ldots, x_n\}$ is a subset of p - dimensional space, here n is number of elements in set

X, then the clustering of X is the partitioning of X into k clusters $\{C_1, C_2 \ldots, C_k\}$ satisfying the following conditions :

(1) $\bigcup_{i=1}^{k} C_i = X$
(2) $C_i \neq \phi \ \forall \ i = 1, 2, \ldots, k$
(3) $C_i \cap C_j = \phi$ where $i \neq j$

Definition 2.2 (Euclidean Distance)**.** [17] Let $X = \{x_1, x_2, \ldots, x_n\}$ is a subset of p - dimensional space, here n is number of elements in set X. Euclidean distance defined as

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2} = \|x_i - x_j\|$$

3. Partitional Clustering Techniques

Algorithms for partitional clustering are typically iterative and reach local maxima. The phases of an iterative clustering method used by Hamerly and Elkan [7] for any data set X are:

(1) Take K cluster centroids randomly
say $\{m_1, m_2, \ldots, m_k\}$ are k cluster centroids.
(2) For each element $x_i \in X$ compute its membership $u(m_j|x_i)$ to each centroid m_k and its weight $w(x_i)$
(3) Recalculate the k cluster centroids, using

$$m_j = \frac{\sum_{\forall x_i} u(m_j|x_i) w(x_i) x_i}{\sum_{\forall x_i} u(m_j|x_i) w(x_i)} \tag{1}$$

Repeat this process up until a stopping requirement is met.
The membership function $u(m_j|x_i)$ in the aforementioned procedure quantifies the membership of pattern x_i to cluster k. The following restrictions must be fulfilled by the membership function, $u(m_j|x_i)$:

(1) $u(m_j|x_i) \geq 0, \ i = 1, 2, \ldots, n$ and for all specified k
(2) $\sum_{j=1}^{k} u(m_j|x_i) = 1, \ i = 1, 2, \ldots, n$

Crisp clustering algorithms employ a hard membership function (i.e. $u(m_j|x_i) \in \{0, 1\}$), whereas fuzzy clustering algorithms employ a soft membership function (i.e. $u(m_j|x_i) \in [0, 1]$)[7].

In Eq. 1, the weight function, $w(z_i)$, defines the amount of influence pattern z_i has on recalculating the centroids in the next iteration, where $w(z_i) > 0$ [7]. Zhang [25] proposed the weight function.

An iterative clustering method may use several stopping criteria, such as:

- Stop when the centroid values have changed by less than a user-specified amount,
- The quantization error is sufficiently modest, or
- When an allotted number of iterations has been reached, stop.

Popular iterative clustering algorithms are described in the following by defining the membership and weight functions in eq. 1.

3.1. The K-Means Algorithm. The K-Means clustering method is one of the most important, extensively studied, and applied algorithms [15]. Its popularity is primarily due to how simple it is to interpret the data. This algorithm divides a set of n items into $k \geq 2$ clusters in an iterative manner. As a result, the objects within a cluster are similar to one another and distinct from those inside other clusters [18]. The K-means algorithm optimizes the objective function :

$$J_{K-means} = \sum_{j=1}^{k} \sum_{\forall x_i \in C_j} d(x_i, m_j) \tag{2}$$

As a result, the K-Means algorithm reduces the intra-cluster distance [7]. The K-Means algorithm begins with K centroids (the initial values for the centroids are chosen at random or derived from prior knowledge). The closest cluster is then assigned to each pattern in the data set (i.e. closest centroid). The centroids are then recalculated in accordance with the associated patterns. This procedure is carried out until convergence is achieved.

K-Means membership and weight functions are defined as

$$u(m_j|x_i) = \begin{cases} 1 & \text{if } d(x_i, m_j) = \arg\min_j \{d(x_i, m_j)\} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$w(x_i) = 1 \tag{4}$$

K-Means has a hard membership function. Additionally, K-Means provides a constant weight function, giving all patterns the same weight.

3.2. 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:

$$J_{FCM} = \sum_{j=1}^{k} \sum_{i=1}^{p} u_{j,i}^{q} d(x_i, m_j) \tag{5}$$

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) $u_{j,i} \geq 0,\ i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, k$
(2) $\sum_{j=1}^{k} u_{j,i},\ i = 1, 2, \ldots, p$

For FCM the membership function is denifend as

$$u(m_j|x_i) = \frac{\|x_i - m_j\|^{-2/(q-1)}}{\sum_{j=1}^{k} \|x_i - m_j\|^{-2/(q-1)}} \tag{6}$$

and weight function is defined as

$$w(x_i) = 1 \tag{7}$$

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 [10].

4. Implementation Methodology

Python's NumPy rand function is used to create a sample 2D array, and the K- Means and FCM default libraries in python are used to process the data. To highlight the differences between the two approaches, we run the procedure for k = 2, 3 and 4 cluster centroids.

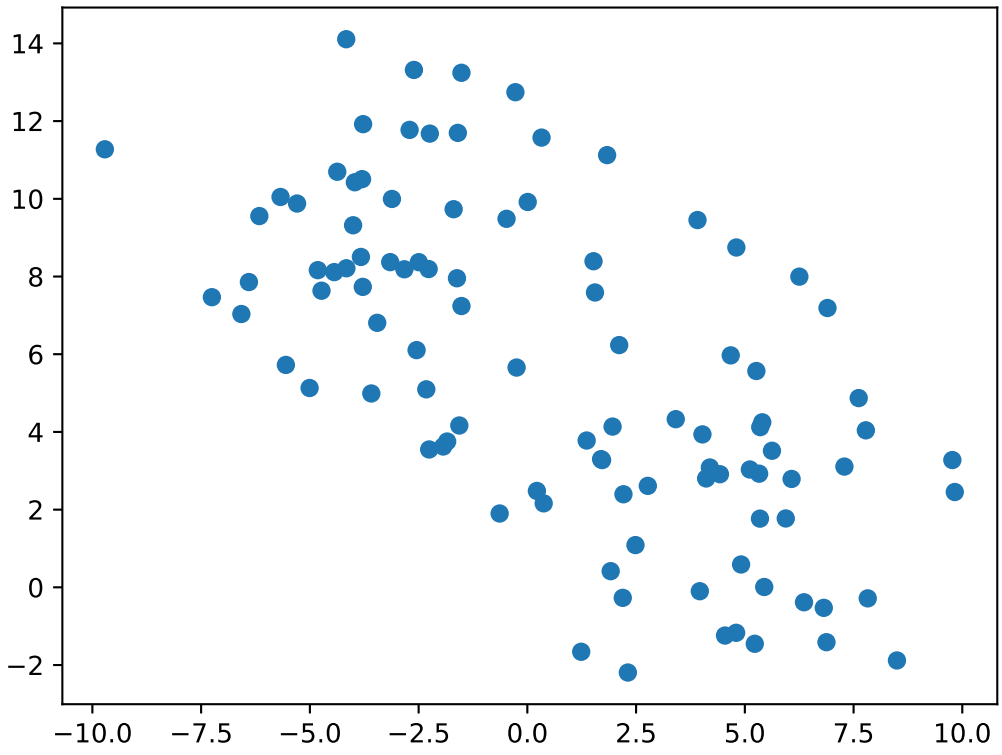
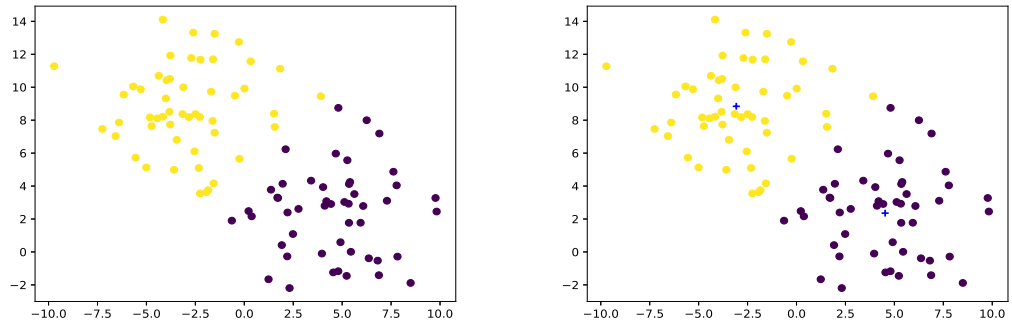


FIGURE 1. Data Set X

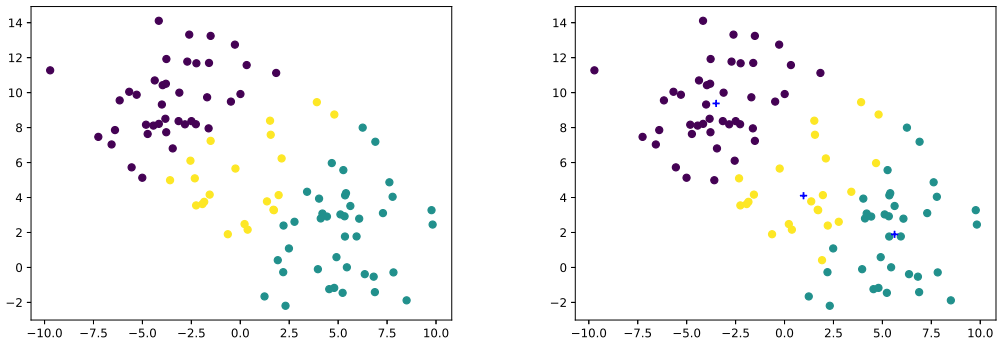


(A) K-Means (B) FCM

FIGURE 2. For $k = 2$ cluster centroids

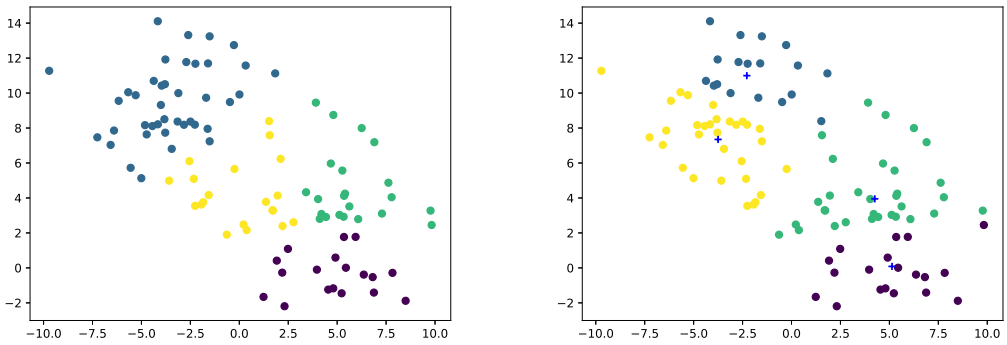
	A	B		A	B		A	B		A	B
1	-3.82693	8.503724	26	1.833638	11.12473	51	4.675634	5.970388	76	5.328734	2.925902
2	7.61827	4.871125	27	4.427849	2.911338	52	-7.25272	7.467995	77	-3.77595	11.92137
3	-4.00625	9.319323	28	-0.27475	12.74395	53	-3.7836	7.733529	78	1.915904	0.415851
4	6.900544	7.18935	29	6.075218	2.789878	54	5.623794	3.515327	79	9.770759	3.27621
5	5.265394	5.567812	30	3.912073	9.453635	55	-1.8438	3.752765	80	7.782642	4.040986
6	-1.51536	13.24389	31	-3.15312	8.370409	56	6.360464	-0.38401	81	3.411963	4.328266
7	5.400779	4.247924	32	-6.57513	7.034715	57	-2.24223	11.67806	82	4.196343	3.08431
8	0.375554	2.161718	33	-5.67443	10.04746	58	5.348142	1.768443	83	4.91392	0.588612
9	-3.5875	4.98962	34	1.521336	8.393401	59	2.193714	-0.27031	84	-2.70723	11.774
10	-4.81705	8.163952	35	-1.56425	4.165926	60	-2.49514	8.369172	85	2.111462	6.235489
11	-2.3235	5.096229	36	6.253736	7.996923	61	-3.1133	9.996346	86	-2.54632	6.105581
12	1.238264	-1.65809	37	5.116126	3.032792	62	2.311196	-2.19266	87	5.445828	0.008703
13	-3.80025	10.50633	38	-1.59822	11.69704	63	0.008158	9.918352	88	-4.16389	14.10805
14	-6.16171	9.555655	39	-4.73256	7.634454	64	0.326051	11.57531	89	-0.2472	5.656966
15	2.769811	2.611867	40	-1.93482	3.625193	65	-3.45166	6.808024	90	-5.00626	5.130451
16	2.485932	1.087501	41	-0.47842	9.485549	66	-1.5147	7.240207	91	1.718544	3.275549
17	4.805512	8.747085	42	6.877051	-1.41171	67	-2.26982	8.192016	92	-2.82723	8.186251
18	1.694924	3.299969	43	1.959504	4.137652	68	4.025356	3.936671	93	-4.43984	8.113215
19	4.799953	-1.17	44	5.941282	1.77289	69	-3.96604	10.42577	94	-9.7135	11.27452
20	-2.25686	3.548472	45	3.965062	-0.09961	70	-0.63674	1.900257	95	4.544421	-1.2407
21	5.228637	-1.45261	46	1.555011	7.589043	71	7.826017	-0.28371	96	-5.29448	9.878466
22	4.110886	2.802425	47	-2.60772	13.31706	72	2.209271	2.395914	97	8.497563	-1.88192
23	6.815217	-0.52765	48	-1.69487	9.732188	73	9.82689	2.453259	98	-4.16095	8.212128
24	7.289163	3.108317	49	-5.55162	5.724718	74	-6.40147	7.857511	99	-1.61797	7.95531
25	5.354557	4.123183	50	1.363794	3.778692	75	0.2193	2.480913	100	-4.37073	10.6964

TABLE 1. Sample 2D array X



(A) K-Means (B) FCM

FIGURE 3. For $k = 3$ cluster centroids



(A) K-Means (B) FCM

FIGURE 4. For $k = 4$ cluster centroids

5. CONCLUSION

In this study, we use the fuzzy C-Mean and K-Means algorithms to a 2D array with variable cluster centroids. We note no difference between FCM and K-Means results for k = 2 cluster centroids (see fig. 2). Both techniques have a small difference for $k = 3$ cluster centroids (see fig. 3). However, both approaches provide a different clustering (partitioning) set for the data set when applied to $k = 4$ cluster centroids (see fig. 4). Therefore, K-Means is advised for cluster centroids with $k = 2$ *or* 3, however FCM provides more accurate findings for cluster centroids with $k > 3$. We have discussed the quality of this cluster algorithms (see [22]).

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

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

[2] R. Bellman, R. Kalaba, L. Zadeh, Abstraction and pattern classification, J. Math. Anal. Appl. 13 (1966), 1-7. https://doi.org/10.1016/0022-247x(66)90071-0.

[3] J.C. Bezdek, A convergence theorem for the fuzzy ISODATA clustering algorithms, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-2 (1980), 1-8. https://doi.org/10.1109/tpami.1980.4766964.

[4] J.C. Bezdek, Objective function clustering, in: pattern recognition with fuzzy objective function algorithms, Springer US, Boston, MA, 1981: pp. 43-93. https://doi.org/10.1007/978-1-4757-0450-1_3.

[5] S. Ever-Hadani, Applications of cluster analysis algorithm to geostatistical series, Region. Sci. Urban Econ. 10 (1980), 123–151. https://doi.org/10.1016/0166-0462(80)90052-6.

[6] S. Ghosh, S.K. Dubey, Comparative analysis of K-means and fuzzy C-means algorithms, Int. J. Adv. Computer Sci. Appl. 4 (2013), 35-39. https://doi.org/10.14569/IJACSA.2013.040406

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

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

[9] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybernetics, 3 (1973), 32-57. https://doi.org/10.1080/01969727308546046.

[10] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: A review, ACM Comput. Surv. 31 (1999), 264-323. https://doi.org/10.1145/331499.331504.

[11] G.M. Lee, X. Gao, A hybrid approach combining fuzzy c-Means-Based genetic algorithm and machine learning for predicting job cycle times for semiconductor manufacturing, Appl. Sci. 11 (2021), 7428. https://doi.org/10.3390/app11167428.

[12] S.J. Lee, D.H. Song, K.B. Kim, et al. Efficient fuzzy image stretching for automatic ganglion cyst extraction using fuzzy C-means quantization, Appl. Sci. 11 (2021), 12094. https://doi.org/10.3390/app112412094.

[13] A. Liew, S. Leung, W. Lau, Fuzzy image clustering incorporating spatial continuity. IEE Proc.-Vis. Image Signal Process. 147 (2000), 185-192.

[14] J. MacQueen, Classification and analysis of multivariate observations, in: 5th Berkeley Symp. Math. Statist. Probability, (1967), 281–297.

[15] M.C. Naldi, R.J. Campello, Comparison of distributed evolutionary k-means clustering algorithms, Neurocomputing, 163 (2015), 78-93.

[16] J. Nayak, B. Naik, H.S. Behera, Fuzzy C-Means (FCM) clustering algorithm: A decade review from 2000 to 2014, in: L.C. Jain, H.S. Behera, J.K. Mandal, D.P. Mohapatra (Eds.), Computational Intelligence in Data Mining - Volume 2, Springer India, New Delhi, 2015: pp. 133-149. https://doi.org/10.1007/978-81-322-2208-8_14.

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

[18] J. Pérez-Ortega, N.N. Almanza-Ortega, D. Romero, Balancing effort and benefit of K-means clustering algorithms in Big Data realms, PLoS One, 13 (2018), e0201874.

[19] E.H. Ruspini, A new approach to clustering, Inform. Control, 15 (1969), 22-32. https://doi.org/10.1016/S0019-9958(69)90591-9.

[20] E.H. Ruspini, J.C. Bezdek, J.M. Keller, Fuzzy clustering: A historical perspective, IEEE Comput. Intell. Mag. 14 (2019), 45-55. https://doi.org/10.1109/MCI.2018.2881643

[21] A.S. Shirkhorshidi, S. Aghabozorgi, T.Y. Wah, et al. Big Data Clustering: A Review, in: B. Murgante, S. Misra, A.M.A.C. Rocha, et al. (Eds.), Computational Science and Its Applications - ICCSA 2014, Springer International Publishing, Cham, 2014: pp. 707-720. https://doi.org/10.1007/978-3-319-09156-3_49.

[22] R.K. Verma, R. Tiwari, P.S. Thakur, Partition coefficient and partition entropy in fuzzy c means clustering, J. Sci. Res. Rep. 29 (2023), 1-6. https://doi.org/10.9734/jsrr/2023/v29i121812.

[23] M.S. Yang, A survey of fuzzy clustering, Math. Computer Model. 18 (1993), 1-16. https://doi.org/10.1016/0895-7177(93)90202-A.

[24] L.A. Zadeh, Fuzzy sets, Inform. Control, 8 (1965), 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X.

[25] B. Zhang, Generalized k-harmonic means - boosting in unsupervised learning, Technical Report HPL-2000-137, Hewlett-Packard Labs, (2000).