# Fuzzy *c*-Means Clustering Strategies: A Review of Distance Measures


**Jyoti Arora, Kiran Khatter and Meena Tushir**

**Abstract** In the process of clustering, our attention is to find out basic procedures that measures the degree of association between the variables. Many clustering methods use distance measures to find similarity or dissimilarity between any pair of objects. The fuzzy *c*-means clustering algorithm is one of the most widely used clustering techniques which uses Euclidean distance metrics as a similarity measurement. The choice of distance metrics should differ with the data and how the measure of their comparison is done. The main objective of this paper is to present mathematical description of different distance metrics which can be acquired with different clustering algorithm and comparing their performance using the number of iterations used in computing the objective function, the misclassification of the datum in the cluster, and error between ideal cluster center location and observed center location.

**Keywords** FCM clustering · Euclidean distance · Standard euclidean distance Mahalanobis distance · Minkowski distance · Chebyshev distance

J. Arora (✉)
Department of Information Technology, Maharaja Surajmal Institute of Technology, C-4, Janakpuri, New Delhi, India
e-mail: joy.arora@gmail.com

K. Khatter
Department of Computer Science, Ansal University, Gurgaon, India
e-mail: kirankhatter@ansaluniversity.edu.in

M. Tushir
Department of Electrical & Electronics Engineering, Maharaja Surajmal Institute of Technology, C-4, Janakpuri, New Delhi, India
e-mail: meenatushir@yahoo.com



M. N. Hoda et al. (eds.), *Software Engineering*, Advances in Intelligent Systems and Computing 731, https://doi.org/10.1007/978-981-10-8848-3_15

# 1 Introduction

Clustering is a technique of finding similar characteristic data among the given set of data through association rules and classification rules resulting into separation of classes and frequent pattern recognition. Clustering is basically knowledge discovery process whose result can be used for future use, in various applications. A good cluster definition involves low interclass similarity and high intra-class similarity. In order to categorize the data, we have to apply different similarity measure techniques to establish a relation between the patterns which will group the data into different clusters with a degree of membership. In clustering, we have to evaluate a good distance metrics, in order to have high intra-class similarity. Several clustering algorithms with the different distance metrics have been developed in the past, some of them are used in detecting different shapes of clusters such as spherical [1], elliptical [2], some of them are used to detect the straight lines [3, 4], algorithms focusing on the compactness of the clusters [2, 5]. Clustering is a challenging field of research as it can be used as a separate tool to gain insight into the allocation of data, to observe the characteristic feature of each cluster, and to spotlight on a particular set of clusters for more analysis. Focusing on the proximity measures, we can find some work that compares a set of distance metrics; therefore, these could be used as guidelines. However, most of the work includes basic distance metrics as Grabusts [6] compared Euclidean distance, Manhattan distance, and Correlation distance with k-means on Iris dataset, similarly Hathaway [7] compared distance with different values of $p$, Liu et al. [8] proposed a new algorithm while changing the Euclidean distance with Standard Mahalanobis distance.

In this paper, we are presenting the survey of different distance metrics in order to acquire proximity measure to be followed by the clustering criterion that results in the definition of a good clustering scheme for dataset. We have included Euclidean distance, Standard Euclidean distance, Mahalanobis distance, Standard Mahalanobis distance, Minkowski distance, and Chebyshev distance and compared on the criteria of accuracy and misclassification, location of center which to our knowledge have not been discussed in such detail in any of the surveys till now.

The remainder of this paper is sectioned as follows. Section 2 provides related work which includes detail of fuzzy $c$-means algorithm and overview of different distance metrics, and Sect. 3 includes experimental results on different data types including synthetic and real datasets. Section 4 concludes the review.

# 2 Related Work

## *2.1 Fuzzy* c*-Means Algorithm*

The notion of fuzzy sets was developed by Zadeh [9] is an attempt to modify exclusive clustering on the basis of their probability of any parameter on which clusters have been developed, which was further extended by Bezdek et al. [3] as

fuzzy $c$-means algorithm (FCM) is the most widely used algorithm. This approach partitions a set of data $\{x_1, \ldots, x_n\} \subset R^s$ into $c$-clusters based on a similarity computed by the least square function of Euclidean distance metrics. The objective function of FCM is

$$J_{\mathrm{FCM}}(X : U, V) = \sum_{j=1}^{c} \sum_{i=1}^{N} (u_{ij})^m \left\|x_i - v_j\right\|^2, 1 < m < \infty \tag{1}$$

where $m > 1$ is a fuzzification parameter, $v_j \in R^S$ is a cluster center, $u_{ij} \in [0, 1]$ is a degree to which data $x$ belongs to cluster, defines partition matrix. $\left\|x_i - v_j\right\|^2$ is a Euclidean distance metrics. The partition matrix $u_{ij}$ is a convenient tool for representing cluster structure in the data, where the fuzzy partition has constraint $\sum_{i=1}^{c} u_{ij} = 1$. The most effective method of optimization of Eq. (1) is by alternative optimization method used in conventional FCM. The equation of $v_j$ and $u_{ij}$ can be referred from [3].

## 2.2 Distance Measures

Important component of clustering algorithm is similarity measurement between the data points. There are different measures which have been used with different clustering algorithms in order to have clusters with desired properties. Some distance measures we will discuss are

### 2.2.1 Euclidean Distance

The Euclidean distance is the most intense similarity measure which is used widely in FCM. This formula includes two objects and compares each attribute of individual item with other to determine strength of relativity with each other. The smaller the distance is greater the similarity. The equation for Euclidean distance is

$$d_{x,v} = \overline{)(x_1 - v_i)^2 + (x_2 - v_i)^2 \ldots (x_n - v_i)^2} \tag{2}$$

Euclidean distance as a measure of similarity, hyperspherical-shaped clusters of equal size are usually detected [10]. However, Euclidean distance degrades the performance in the presence of noise in the dataset. This is because the object to center dissimilarity term in (1) can place considerable membership to the outlying data and due to membership constraint.

### 2.2.2 Standard Euclidean Distance

The Standard Euclidean distance can be squared in order to place progressively greater weight on objects that are farther apart. In this case, the equation becomes

$$d_{x,v}^2 = (x_1 - v_i)^2 + (x_2 - v_i)^2 \ldots (x_n - v_i)^2 \tag{3}$$

This distance metrics is frequently used in optimization problems in which distances only have to be compared. Clustering with the Euclidean Squared distance is faster than clustering with the regular Euclidean distance.

### 2.2.3 Mahalanobis Distance

Mahalanobis distance is a measure of the distance between a given point and a distribution. It is a multi-dimensional generalization of the act of determining the number of standard deviations that a point $x$ is away from the mean of the distribution. The equation of the Mahalanobis distance is given by

$$d_{x,v} = \overline{)(x_k - v_i)A^{-1}(x_k - v_i)} \tag{4}$$

In (4), A is a covariance matrix of data. The Mahalanobis distance is better adopted where the cluster required are nonspherical in shape. In [2], Cai has discussed in clustering algorithm, a modified Mahalanobis distance with preserved volume was used. It is more particularly useful when multinormal distributions are involved.

### 2.2.4 Standard Mahalanobis Distance

In Standard Mahalanobis distance, covariance matrix is replaced by correlation matrix. The equation of the Standard Mahalanobis distance is represented by

$$d_{x,v} = \overline{)(x_k - v_i)R^{-1}(x_k - v_i)} \tag{5}$$

In (5) R is a correlation matrix. In [8], Liu et al. has proposed new algorithm giving FCM-SM, normalizing each feature in the objective function, and all covariance matrix becomes corresponding correlation matrix.

### 2.2.5 Minkowski Distance

Minkowski distance termed as $L_p$ is a generalized metric that includes special cases of $p$ and introduced three distance metrics with $p = 1$(Manhattan distance), $p = 2$

(Euclidean distance), and $p = \infty$ (Chebyshev distance). The Minkowski distance of order $p$ between data points and center points is represented by

$$d_{x,y} = \left( \sum_{i=1}^{n} \|x_i - v_i\|^p \right)^{1/p} \tag{6}$$

In [1, 2], different function of Minkowski distance with different value of $p$ has been implemented with FCM showing results on relational and object data types.

#### 2.2.6 Chebyshev Distance

Chebyshev distance ($L_\infty$) is a distance metric specified on a vector space where the given two vectors are separated by a distance which is the largest of their differences measured along any coordinate dimension. Essentially, it is the maximum distance between two points in any single dimension. The Chebyshev distance between ant two points is given by (7)

$$d_{x,y} = \max|x_i - v_i| \tag{7}$$

## 3 Experiments' Results

The purpose of the experimental part was to test the operation of the FCM algorithm by applying different distance metrics on synthetic and real dataset. We used datasets with wide variety in the shape of clusters, numbers of clusters, and count of features in different data point. FCM is an unsupervised clustering so the number of clusters to group the data was given by us. We choose $m = 2$ which is a good choice for fuzzy clustering. For all parameters, we use $\varepsilon = 0.001$, max_iter = 200.

### *3.1 Synthetic Datasets*

#### 3.1.1 $X_{12}$ Dataset

The first example involves $X_{12}$ dataset as given in Fig. 1 contains two identical clusters with one outlier which is equidistant from both the clusters. We know FCM is very sensitive to noise, so we do not get the desired results. To show the effectiveness of different distance metrics, we have calculated the error $E_*$, by sum of the square of the difference between calculated center and the ideal center with every distance metrics as given in Eq. (8).

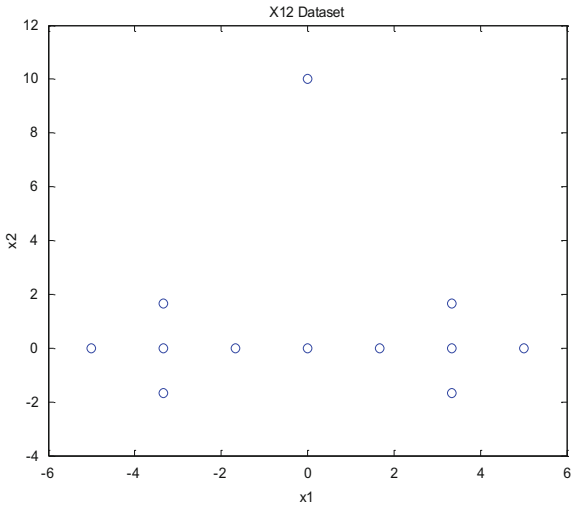


**Fig. 1** Representation of $X_{12}$ dataset

**Different Volume Rectangle Dataset**

By changing the volume of clusters in a pattern, we observe the effectiveness of different distance metrics. Figure 2 shows the representation of dataset with two identical clusters of different volume, ideal cluster center are defined by $V_{\mathrm{RD}}$.

Table 1 shows that for $X_{12}$ dataset, best results are shown with Euclidean distance, Minkowski distance, and Chebyshev distance but number of iterations used by Minkowski distance are more. The Standard Mahalanobis distance and

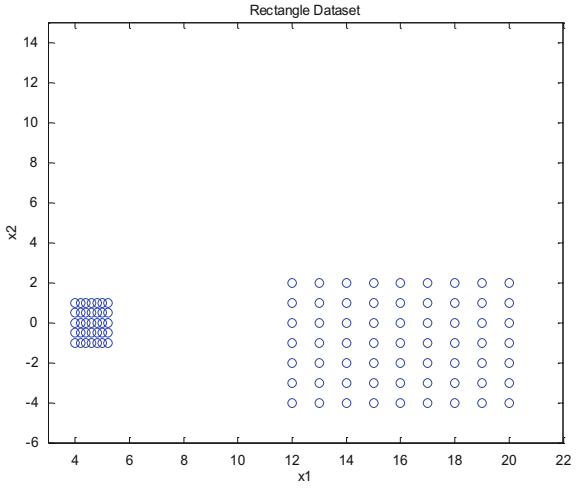


**Fig. 2** Representation of rectangle dataset

**Table 1** Centers produced by FCM with different distance metrics, effectiveness and number of iteration for the $X_{12}$ dataset and different volume rectangle dataset

| Distance | Euclidean | | Std. Euclidean | | Mahalanobis | | Std. mahalanobis | | Minkowski | | Chebyshev | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Center ($V_{12}$) | 2.98<br>−2.98 | 0.54<br>0.54 | 2.67<br>−1.51 | 2.15<br>0.08 | 2.97<br>−2.97 | 0.55<br>0.55 | −2.98<br>2.98 | 0.54<br>0.54 | −2.98<br>2.98 | 0.54<br>0.54 | −2.98<br>2.98 | 0.54<br>0.54 |
| $E_*$ | 0.412 | | 4.213 | | 0.438 | | 0.412 | | 0.412 | | 0.412 | |
| No. of Iter. | 10 | | 10 | | 10 | | 9 | | 17 | | 10 | |
| Center ($V_{RD}$) | 4.22<br>16.30 | −0.02<br>−1.01 | 7.42<br>16.26 | 0.30<br>−2.02 | 5.90<br>16.42 | −0.10<br>−1.06 | 4.86<br>16.33 | 0.54<br>−0.94 | 4.92<br>16.33 | −0.01<br>−0.94 | 4.76<br>16.26 | −0.02<br>−1.00 |
| $E_*$ | 0.349 | | 3.53 | | 0.5 | | 0.211 | | 0.059 | | 0.062 | |
| No. of Iter. | 10 | | 23 | | 27 | | 10 | | 14 | | 13 | |

Mahalanobis distance also show optimal results but Standard Euclidean distance shows poor results. Similarly in Different Volume Rectangle dataset, Minkowski distance and Chebyshev distance perform best as compared to other distance metrics. The highest number of iterations is used by Mahalanobis distance. The Standard Euclidean distance shows worst result with this dataset also. Both the above data comprises of clusters forming compact clouds that are well separated from one another, thus sum of squared error distance outperforms as compare to other distance and Standard Euclidean distance shows poor results. Mahalanobis distance due to calculation of covariance matrix for the data does not show accurate result.

$$E_* = \|V_{12}(\text{ideal}) - V_{12}(\text{dist})\|^2 \tag{8}$$

$$V_{12}(\text{ideal}) = \begin{bmatrix} -3.3400 & 0 \\ 3.3400 & 0 \end{bmatrix} \quad V_{\text{RD}}(\text{ideal}) = \begin{bmatrix} 5 & 0 \\ 16 & -1 \end{bmatrix}$$

## 3.2 High-Dimensional DataSets

We now examine the defined evaluation criteria with some well-known real datasets, namely Iris dataset, Wine dataset and Wisconsin dataset. We are going to analyze the clustering results using Huang' s accuracy measure ($r$) [11].

$$r = \frac{\sum_{i=1}^{k} n_i}{n} \tag{9}$$

where $n_i$ is the number of data occurring in both the $i$th cluster and its corresponding true cluster, and $n$ is the number of data points in the dataset. According to this measure, a higher value of r indicates a better clustering result with perfect clustering yielding a value of $r = 1$.

We made several runs of FCM with different distance metrics and calculated the misclassification, accuracy, number of iterations on all the three high-dimensional datasets. Here in Table 2, we find that how the algorithm shows different values of misclassification over the three datasets with the change of distance metrics. In this, we can see Chebyshev distance is giving good result with Iris and Breast Cancer dataset with an accuracy of 90 and 96% respectively, Standard Euclidean distance is giving best result with Wine dataset with an accuracy of 91%, however number of iterations used is very high.

**Table 2** FCM with different distance metrics showing misclassification, accuracy and number of iteration with Iris, Wine, Breast Cancer (BC) dataset

| Dataset | Distance | | | | | |
|---|---|---|---|---|---|---|
| | Euclidean | Std. Euclidean | Mahalanobis | Mahalanobis | Minkowski | Chebyshev |
| $\text{Misclassification}_{\text{IRIS}}$ | 17 | 27 | 43 | 27 | 17 | 15 |
| $\text{Accuracy}_{\text{IRIS}}$ | 0.88 | 0.82 | 0.71 | 0.82 | 0.88 | 0.9 |
| $\text{No. of Iter.}_{\text{IRIS}}$ | 18 | 41 | 67 | 22 | 25 | 20 |
| $\text{Misclassification}_{\text{WINE}}$ | 56 | 16 | 79 | 55 | 55 | 56 |
| $\text{Accuracy}_{\text{WINE}}$ | 0.68 | 0.91 | 0.55 | 0.69 | 0.69 | 0.68 |
| $\text{No. of Iter.}_{\text{WINE}}$ | 47 | 113 | 9 | 46 | 86 | 61 |
| $\text{Misclassification}_{\text{BC}}$ | 30 | 31 | 22 | 53 | 38 | 22 |
| $\text{Accuracy}_{\text{BC}}$ | 0.95 | 0.95 | 0.96 | 0.92 | 0.94 | 0.96 |
| $\text{No. of Iter.}_{\text{BC}}$ | 14 | 22 | 100 | 25 | 21 | 18 |

## 4 Conclusions

We described various distance metrics for FCM and examined the behavior of the algorithm with different approaches. It has been concluded from results on various synthetic and real datasets that Euclidean distance works well for most of the datasets. Chebyshev and Minkowski distances are equally suitable for clustering. Further exhaustive exploration on distance metrics needs to be done on various datasets.

## References

1. Patrick, J.F., Groenen, U., Kaymak, J.V., Rosmalen: Fuzzy clustering with Minkowski distance functions. Econometric Institute Report, **EI**(24), (2006)
2. Cai, J.Y., Xie, F.D., Zhang, Y.: Fuzzy c-means algorithm based on adaptive Mahalanobis distance. Comput. Eng. Appl. 174–176(2010)
3. Bezdek, J.C., Coray, C., Gunderson, R, Watson, J.: Detection and characterization of cluster substructure. SIAM J. Appl. Math. 339–372 (1981)
4. Dave, R.N.: Use of the adaptive fuzzy clustering algorithm to detect lines in digital images. Intell Robots Comput. Vision VIII 1192, pp. 600–661 (1982)
5. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. J Cybern, 32–57(1973)
6. Grabusts, P.: The choice of metrics for clustering algorithms. In: International Scientific and Practical Conference Vol 2(8), pp. 70–76 (2011)
7. Hathaway, R.J., Bezdek, J.C., Hu, Y.: Generalised fuzzy c-means clustering strategies using $L_P$ norm distance. IEEE Trans. Fuzzy Syst. **8**(5), (2000)
8. Liu, H.C., Jeng, B.C., Yih, J.M., Yu, Y.K.: Fuzzy c-means clustering algorithm based on standard mahalanobis distance. In: International Symposium on Information Processing, pp. 422–427 (2009)
9. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
10. Su, M.C., Chou, C.H.: A Modified means of K-Means algorithm with a distance based on cluster symmetry. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 674–680 (2001)
11. Tushir, M., Srivastava, S.: A new kernelized hybrid c-mean clustering with optimized parameters. Appl. Soft Comp. **10**, 381–389 (2010)