Clustering tourism places in Madura based facilities using fuzzy C-means

Cite as: AIP Conference Proceedings **2679**, 020007 (2023); https://doi.org/10.1063/5.0111264
Published Online: 04 January 2023

Eka Mala Sari Rochman, Sri Herawati, Ach. Khozaimi, et al.

Clustering Tourism Places in Madura Based Facilities Using Fuzzy C-Means

Eka Mala Sari Rochman[1,a)], Sri Herawati[1], Ach. Khozaimi[1], Endang Indriyani[1], Bain Khusnul Khatimah[1], Aeri Rachmad[1]

[1]*Department of Informatics, Faculty of Engineering, University of Trunojoyo, Madura, Bangkalan, Indonesia*

[a)] *Corresponding author: ekamalasari3@gmail.com*

Abstract. Madura Island consists of four districts namely Bangkalan, Sampang, Pamekasan, and Sumenep. Each district has many choices of interesting tourist attractions. There are various types of tourism ranging from natural tourism, cultural tourism, historical tourism, and artificial tourism. With the diversity of these tourist attractions, it is enough to invite many tourists to come on vacation to Madura Island. Each tourist attraction has a different number of visitors, the number of public facilities, and ticket prices. All the criteria possessed by each tourist attraction have their own assessment for potential tourists. The local Tourism Office must know the developments in each tourist attraction, so that it can maintain the quality of the tourist attraction. Improvement of infrastructure can be through public facilities provided at tourist objects is very necessary. The purpose of this study was to determine how well the Fuzzy C-Means method in grouping tourism objects in Madura. Fuzzy C-Means is a grouping method which the development of the k-means cluster is not a hierarchical method that allocates data into each group by utilizing fuzzy set theory. From the trials that have been carried out, the best grouping results are found in cluster 10 with a silhouette coefficient value of 0.825 which is included in the strong category.

INTRODUCTION

Tourism is a human activity in making a temporary trip from their place of residence, to one or several destinations outside the environment of residence without intending to earn a living [1][2]. Tourism is one of the sectors driving the economy that needs to be given more attention in order to develop properly. Madura Island is one of the islands in Indonesia which has a lot of tourism potential in each district. There are four districts on the island of Madura, namely Bangkalan, Sampang, Pamekasan and Sumenep. The available public facilities at tourist objects are the important factors that can attract tourists to come and visit these attractions.

Many tourists choose Madura Island as a vacation destination, both domestic and foreign tourists. The increase in tourists visiting Madura is currently a challenge in itself to contribute so that data can be processed into more useful information [3]. One of the uses of data that can be applied is the grouping of tourist objects. In grouping this, it can be done by using several parameters such as the number of visitors, the number of public facilities, ticket prices, and categories of tourist objects.

In this case, the application of data mining can be a solution by analysing the data. Please note that data mining is a tool that allows to access large amounts of data in a fast time. One way of processing data in data mining is Clustering. Clustering is the process of dividing data in a set into several groups that have data similarity greater than other groups [2][4].

The First International Conference on Neuroscience and Learning Technology (ICONSATIN 2021)
AIP Conf. Proc. 2679, 020007-1–020007-6; https://doi.org/10.1063/5.0111264
Published by AIP Publishing. 978-0-7354-4300-6/$30.00

The method that has been used to classify tourist objects is the K-Means and Fuzzy C-Means (FCM) methods. Compared to the FCM method, the K-Means method is more widely used in classifying data [5]. Clustering techniques are generally applied to quantitative or qualitative (categorical) data [6].
The FCM algorithm is used for data sets with many (varied) attributes, while K-Means is more used for datasets with few attributes. Fuzzy C-means works by identifying outliers based on the density of the data set and then grouping them by considering the actual data points. Thus, it can improve the accuracy of the clustering. In this study build a grouping system regarding tourism based on public facilities by applying the fuzzy C-means method.

METHODS

Tourist

Tourism can be interpreted as an activity carried out by a person to go to a tourist place outside of his daily life and environment to make a temporary stopover to get pleasure [7]. In the economic field, tourism is one sector that needs to be given more attention in order to develop properly because information about tourist areas is very easy to obtain and attracts tourists [3].

The development of tourism in an area cannot be separated from the facilities contained therein. Public facilities are facilities and infrastructure where the manager must provide for the interests and convenience of the visitors. Tourism facilities and infrastructure are the main needs in the tourism sector, in this case the visitors will feel all their needs are met if the services and facilities obtained are in accordance with their needs. Some of the facilities that must be available in every tourist attraction are toilets, places of worship, relaxing seats to rest, comfortable places to eat, adequate parking, trash cans so that cleanliness is always maintained, a gift shop, as well as officers who guard the place. the tour so that visitors become safer. Several types of tourism that exist today include Cultural Tourism, Maritime or Maritime Tourism, Nature Reserve Tourism, Convention Tourism, Hunting Tours, Pilgrimage Tours [8].

Clustering

Clustering potential is used to determine the structure in the data that can be used further in a wide variety of applications such as classification, image processing, and pattern recognition [6]. The process of dividing data into subsets based on predetermined similarity or similarity is a method of cluster analysis. So, it can be said in general that data that has a high level of similarity is in the same cluster, which has a low level of similarity will be in a different cluster [1].

Fuzzy

Fuzzy is linguistically defined as fuzzy or vague. A value can be large or false at the same time. The degree of membership in the fuzzy method has a value range of 0 (zero) to 1 (one) [8][9]. This is of course different from the firm set which has a value of 1 or 0 (yes or no). Logic that has a value of fuzziness between true or false is the meaning of fuzzy logic. In fuzzy logic theory, a value can be true or false together. But how big the presence and error depend on the weight of the membership it has.

Fuzzy C-Means (FCM)

Fuzzy C-Means is a data grouping method in which the degree of membership determines the existence of each data point in a cluster [10]. FCM uses a fuzzy grouping model with a fuzzy index using Euclidean Distance, so that data can be members of all classes or clusters formed with different membership degrees between 0 to 1 [5]. The basic concept of FCM, first is to determine the center of the cluster, which will mark the average location for each cluster. In the initial conditions, the cluster center is still not accurate. Each data point has a degree of membership for each cluster that is formed. By fixing the center of the cluster and the degree of membership of each data point repeatedly, it can be seen that the center of the cluster shift to the right location. This iteration is based on the minimization of the objective function that describes the distance from a given data point to the center of the cluster which is weighted by the degree of membership of the data point. The Fuzzy C-Means algorithm is as follows [5][11]:

1. Enter the data to be clustered into a matrix X, where the matrix is m x n, where m is the number of data to be clustered and n is the attribute of each data. Example X_{ij} = i-th data (i=1, 2…m), j-th attribute (j=1, 2…n).
2. Determine:
    a. Number of clusters = c;
    b. Rank/weight = w;
    c. Maximum iteration = MaxIter;
    d. Expected error = ξ;
    e. Initial Objective Function = P0 = 0;
    f. Early iteration;
3. Generate random numbers μ_{ik} (with i=1, 2…m and k=1, 2…c) as elements of the initial partition matrix U, where Xi is the i-th data

$$U = \begin{bmatrix} \mu_{11}(X_1) & \mu_{21}(X_1) \dots . \mu_{c1}(X_1) \\ \mu_{1i}(X_i) & \mu_{2i}(X_i) \dots \dots \mu_{ci}(X_i) \end{bmatrix} \quad (1)$$

With the number of each column in one row is 1 (one).

$$\sum_{i=1}^{c} \mu_{ci} = 1 \quad (2)$$

4. Calculate the center of the k-th cluster: V_{kj} , with k=1,2,…,c and j = 1,2,…,n

$$V_{kj} = \frac{\sum_{i=1}^{m} (\mu_{ik})^w * X_{ij}}{\sum_{i=1}^{m} (\mu_{ik})^w} \quad (3)$$

Where μw is the membership value raised to the power of weight (w)

5. Calculate the objective function at the t-th iteration, Pt:
6. Calculate the change in the degree of membership of each data in each cluster (fixing the partition matrix U)
7. Check stop conditions:7. Check stop conditions
    a. If (|Pt – Pt-1| < ξ) or (t>MaksIter) then stop;
    b. Otherwise: t = t+1, repeat step 4

Silhouette Coefficient

Testing of the Fuzzy C-Means Algorithm Implementation system can be done by testing the validation using the Silhouette Index by calculating the distance using the Manhattan calculation. Silhouette Coefficient is used to see the quality and strength of the cluster, how well an object is placed in a cluster. This method is a combination of cohesion and separation methods [11].

The stages of calculating the Silhouette Coefficient are as follows [10]:

1. Calculate the average distance of the object with all other objects that are in one cluster with the equation:

$$a(i) = \frac{1}{[A] - 1} \sum j \in_{A, j \neq i} \quad d(i, j) \quad (4)$$

a(i) = Average difference of object (i) to all other objects on A, d(i,j) is distance between data i and data j. and A is a Cluster

2. Calculate the average distance of the object with all other objects that are in other clusters, then take the minimum value with the equation:

$$d(i, C) = \frac{1}{|c|} \sum j \in C \quad d(i, j) \quad (5)$$

d(i,C) = Average difference of object (i) to all other objects in C, with C is cluster other than cluster A

3. After calculating d(i,C) for all X, then the smallest value is taken with the equation:

$$b(i) = min_{c \neq a} di, C \quad (6)$$

Cluster B that reaches the minimum i.e., d(i,B) is called the neighbour of object(i). This is the second-best cluster for object(i).

4. Calculate the value of the Silhouette Coefficient with the equation:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (7)$$

The results of the calculation of the Silhouette Coefficient value vary with a range of -1 to 1. The clustering value can be said to be good if it is positive, namely (ai < bi) and ai is close to 0. With this, the maximum Silhouette Coefficient value will be 1 when ai = 0. If s(i) = 1 indicates that cluster i has been in the right cluster. However, if the value of s(i) is 0 then object i is between two clusters, so the object can be said to have an unclear structure [12].

RESULT

The data used were tourism data from four districts in Madura, namely Bangkalan, Sampang, Pamekasan, and Sumenep, namely from 2015 to 2019. The parameters used for grouping tourism objects in Madura were based on the number of tourist attraction visitors, the number of public facilities, ticket prices and types of attractions

System Design

The first step taken in the calculation process of the Fuzzy C-Means Algorithm was to enter the data to be clustered into an X matrix, then determine the initial values such as the number of clusters, rank or weighting, maximum iteration, smallest expected error, and initial iteration. After that, it generates the objective function or the initial partition matrix. The next was to calculate the cluster center, the objective function, and the change in the partition matrix. If the epsilon value is met then the process is complete, if not the calculation will continue to perform calculations. The system design is set out in Figure 1 below

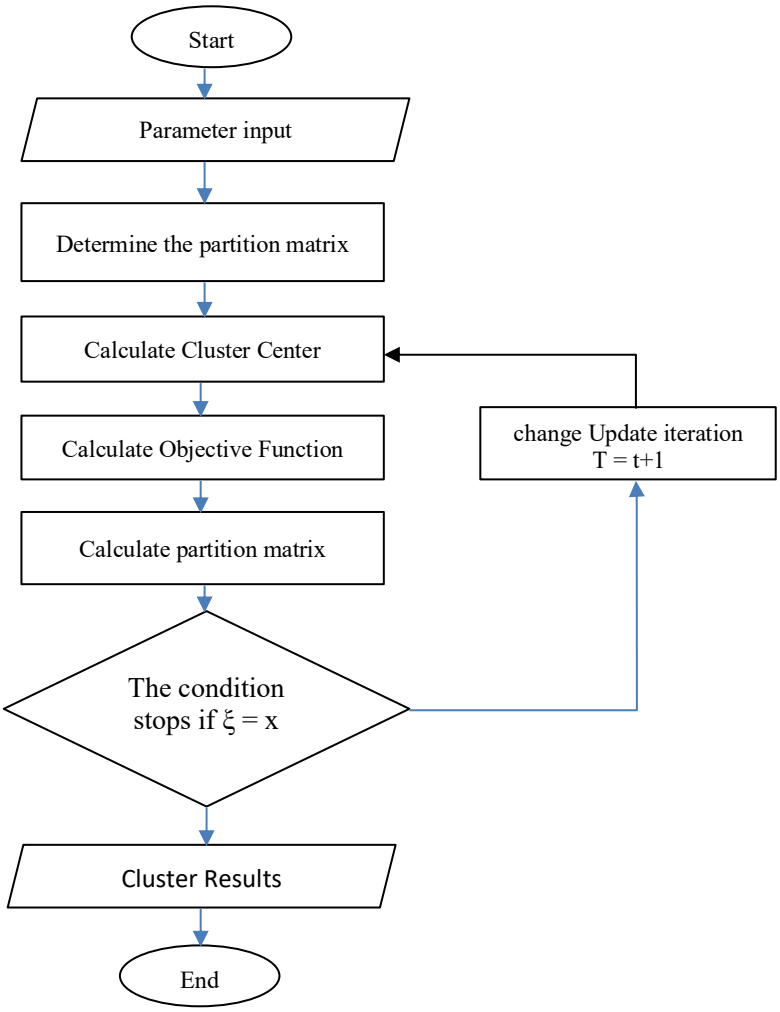

FIGURE 1. Flowchart Fuzzy C-Means

System test

Based on the data and system design, several calculation components for the test scenario are determined which are shown in Table 2 as follows:

TABEL 1. Trial Scenario

No	Calculation Components	Information
1.	Number of clusters	c = 2
2.	Rank (weighting)	w = 2
3.	Maximum Iterations	T = 100
4.	Smallest error	ξ = x
5.	Initial Objective Function	P0 = 0
6.	Early iteration	t = 1

Table 2 explains that: step 1 is initializing by determining the number of clusters to be formed (c = 2), Step-2 determining the weight of the rank (w = 2), because the weight of the rank (w > 1). The third step is to determine the maximum iteration value (T = 100). The 4th step determines the smallest error (e =) which means that it does not exist. Step 5 Objective Function starts from 0 ($P0$ = 0). The 6th step of the initial iteration starts from 1 (t = 1).

TABEL 2 Silhouette Coefficient Values

Number of clusters	Silhoute Coeffisien
2	0,639
3	0,773
4	0,758
5	0,599
6	0,828
7	0,779
8	0,692
9	0,797
10	0,825

Based on the test scenario in Table 2, to test the results of cluster membership in the FCM calculation using the Silhouette Coefficient, this is shown in Table 3, the tests were carried out on Madura tourism objects data based on all criteria, namely the number of visitors, the number of public facilities, ticket prices, and tourist attraction category by entering several central points which are then calculated the silhouette coefficient value from each center point that was entered. Then the best value was in the 10th cluster with a silhouette coefficient value of 0.825, which means that it was included in the strong category

CONCLUSION

Based on research and testing conducted by comparing the silhouette coefficient values from c = 2 to c = 10, it can be concluded that the best performance of the Fuzzy C-Means Method for data on the number of visitors was obtained

when the value of c = 10 reached a silhouette coefficient value of 0.825. It means structure each cluster membership was correct and the resulting cluster is the best cluster.

ACKNOWLEDGMENTS

The author expresses his gratitude to University of Trunojoyo Madura for the opportunity to conduct research in the 2021 Research Group. The author also expresses his gratitude to the Tourism Office throughout Madura for the information regarding tourism data.

REFERENCES

1. Rochman E M S, Rachmad A. *Advances in Social Science, Education and Humanities Research*, volume 410, (2020).
2. Rochman E M S, Pratama I, Husni, Rachmad A. Journal of Physics: Conference Series 1477 022033, (2020)
3. F. Z. d. R. D. Suprihardjo. *JURNAL TEKNIK POMITS* Vol. 3, No.2, 2337-3520, (2014).
4. A. P. Windarto. *Int. J. Artif. Intell. Res.*, vol. 1, no. 2, p. 26, (2017).
5. Gosaina A, Dahiya S., *Procedia Computer Science*, 79, 100 – 111, (2016).
6. Pukhon K K, Hemanta K. Baruah, *International Journal of Cognitive Research in science, engineering and education*, Vol. 1, No.2, (2013).
7. Anamisa, D. R., Rochman, E. M. S., & Rachmad, A., *Journal of Multidisciplinary Engineering Science and Technology (JMEST),* Vol. 6 Issue 1, pp.9446-9448, (2019).
8. Anamisa, D. R., Rachmad, A., & Rochman, E. M. S., *In Journal of Physics: Conference Series*, Vol. 1477, No. 5, p. 052053. IOP Publishing (2020).
9. Anamisa, D. R., Rachmad, A., & Widiastutik, R., *Journal of Theoretical and Applied Information Technology*, 92(1), 52, (2016).
10. Kusumadewi S., Graha Ilmu, Yogyakarta, (2010)
11. Ghosh S, Dubey SK., *International Journal of Advanced Computer Science and Applications*, Vol. 4, No.4, (2013).
12. D. F. Pramesti and C. D. M.Tanzil Furqon. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 1, no. 9, pp. 723-732, (2017)