cek_Optimization Metric Space on Fuzzy C-Means Clustering

by Sugiyarto Sugiyarto

Submission date: 08-May-2021 08:48AM (UTC+0700)

Submission ID: 1581035820

File name: INCEAS_LAST.docx (118.56K)

Word count: 3113

Character count: 16973

Optimization Metric Space on Fuzzy C-Means Clustering

Joko Eliyanto[1] and Sugiyarto Surono[2], Salafudin[3]

[1,2]Department of Mathematics, Ahmad Dahlan Universituny, Yogyakarta, Indonesia
3 Department Of Tadris Mathemathics,FTIK IAIN Pekalongan, Indonesia

Email: [2]sugiyarto@math.uad.ac.id

Abstract. Fuzzy C-Means Clustering is a form of clustering based on distance which apply the concept of fuzzy logic. The clustering process works simultaneously with the iteration process to minimize the objective function. This objective function is an addition from the multiplication of the distance between the data coordinates to the nearest cluster centroid with the degree of which the data belong to the cluster itself. Based on the objective function equation, the value of the objective function will decrease by increasing the number of iteration process. This paper have an objection to observe the value of the distances which usually used whether it will meet the properties of the equation to measure the suitable distance for the application of the Fuzzy C-Means Clustering method. Few numbers of distance are applied on the same set of data. 5 of the standard set of data, and 2 of the random set of data are used to determine the Fuzzy C-Means Clustering performance with 7 different distances. Accuracy, Purity, and Rand Index also being used to measure the result quality of the cluster. The observation shows that the Euclidean distance and Average distance are the distances which provide the optimum result. The suitable distance for the application of the Fuzzy C-Means Clustering method are Euclidean distance, Average distance, Manhattan distance, Minkowski distance, Minkowski-Chebisev distance, and Canberra distance. These six distances are the distances which meet the basic assumption of the Fuzzy C-Means Clustering objective function. The distance which doesn't satisfy the basic assumption is Chebisev distance.

Keywords: Optimization,Fuzzy C Means, Metric Space,

1. Introduction

The Fuzzy Clustering method is a clustering method based on the membership degree of the data [1]. This method allows the membership degree of the data can be considered to certain classification. In Fuzzy Clustering, the method which mostly used is Fuzzy C-Means [2]. The objective function of the Fuzzy C-Means is multiplication value of the membership degree of the data in certain cluster with the square value of the distance difference between the data coordinate to the cluster centroid [3]. There are few variation of the distance that can be used on Fuzzy C-Means clustering. This distance is defined as a function that meet the required criteria [4]. First, the value of distance between two coordinate is non-negative. Second, the value of the distance is zero if and only if this two points are the in the same coordinate. Third, the distance between A to B or B to A is equivalent. Fourth, the distance function has to meet the concept of triangle inequality [5].

The research of the effect of the distance in clustering method has been done a lot. A few of the research conclude with the conclusion where the various function of the distance that being used produces an output which are not much different and there are no sign of the distance that appear dominant [6]–[8]. The effect of the variation of distance function which applied to clustering method, Euclidean distance, Manhattan / City Block distance, Chebisev / Maximum distance, and Minkowski distance, have already identified on K-Means Clustering algorithm [6], [9], [10]. In another research, Euclidean distance, Manhattan/City Block distance, Canberra distance, and Chebisev distance, are applied and evaluated on fuzzy clustering algorithm. The result of that research concluded that the result of the clustering process is very dependent on the set of data that being applied [11], [12].

Apart from the distance function that already mentioned, there are another variation of the distance function which can be applied on clustering algorithm. There are Standardized Euclidean distance, Mahalanobis distance, Cosine distance, Spearman distance, Canberra distance, Bray Curtiz distance, Average distance, Chord distance, Weighted Euclidean distance, Hausdorf distance, and Minkowski-Chebisev distance [7], [13]–[16]. The different application of a certain distance function can increase the performance quality of the clustering algorithm. For example, the combination of the Minkowski and Chebisev distance is proven to improving the algorithm of the Fuzzy C-Means Clustering [15]. The output result then can be improved again by applying the reduction of PCA dimension on the high dimension data [17], [18]. Another way to improve the performance quality of the clustering algorithm is by using Average distance [19], [20].

The main objective from the Fuzzy C-Means Clustering is to minimize the value of the objective function [21]. Intuitively, this method has the quality to improving the clustering result on each iteration by updating the membership degree on every dataset points. The more iteration process is applied, the matrix of the membership degree is able to accurately represent the membership degree of a certain data based on its distance to the existing centroid clusters [22]. Based on that statement, the value of the objective function should be minimized as the number of the iteration is increased. This research focused on the observation to determine the suitable distance to be applied on Fuzzy C-Means Clustering method, where the objective function is tend to decreasing. This result will prove that the Fuzzy C-Means Clustering method works well. The main problem that needs to solve in this research is how to determine which objective function to apply and optimizing the result of the Fuzzy C-Means Clustering method which meet the basic concept of the Fuzzy C-Means Clustering and resulting in a good quality of clustering. The data that being used also limited on 5 UCI machine learning data, which are Hill Valley dataset, Iris dataset, Seeds dataset, and Sonar dataset. This research also provide two another random dataset with certain specification functioned as an addition variable to analyse. The method that being used is developed on the standard Fuzzy C-Means clustering method with limitation on the distance function that being used.

2. Preliminaries

2.1. *Fuzzy C-Means Clustering*

Fuzzy C-Means (FCM) is a form of clustering method which enable one set of data to be classified into two or more clusters [20]. This method is invented by Dunn in 1973 and improved further by Bezdek in 1981. This method is commonly used to introduce certain pattern of data. This method is based on minimalization the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 , \quad 1 \le m < \infty \qquad (1)$$

Where m is any real number valued more than 1, u_{ij} is a membership value of x_i, on cluster j, x_i shows the data in i, c_j is a cluster centroid in j, and $\left\| * \right\|$ is a norm which determine the similarity between data and cluster centroid.

Fuzzy partition is determine by repeated optimization from the objective function which mentioned below with an update of matrix membership u_{ij} and cluster centroid c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{2}$$

and

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{3}$$

is

The iteration process will top when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where "epsilon" is a termination criteria between 0 and 1, where k is the number of iteration process. This procedure is convergent to local minimum value of J_m.

The Fuzzy C-Means Clustering algorithm is consist of the following steps [23]:

a. Initialization of $U = matriks\left[u_{ij}\right], U^{(0)}$

b. In steps of k : measure the vector of the cluster centroid $C^{(k)} = \left[c_j\right]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

c. Updating $U^{(k)}, U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

d. If $\left\|U^{(k+1)} - U^{(k)}\right\| < \varepsilon$, then stops the iteration, if not, back to step 2.

Objective function and membership function are closely related in this method. An assumption where every data coordinates already has their own value of membership degree for each existing cluster is used on the initial process of iteration. Then, this value will be updated over and over by using equation (2). During this process, equation (1) will also be updated to reach its minimum value. Figure 1 illustrate the value distribution of the data membership in the initial of iteration process. Afterwards, the output of the iteration will be updated in every iteration (Figure 2). The iteration process will be terminated when the value of objective function reach their minimum where each data on every cluster has their ideal membership distribution which illustrated on Figure 3. The ideal membership distribution of the data are occur because every data that being analyzed are gathered on their closest cluster centroid..

2.2. Distance Functions

The distance function which used in this research are:

2.2.1. Euclidean Distance

Euclidean distance is known as the most common and standard distance function to apply for clustering method. This function for point x and y is define with this following equation [19]:

$$d_{euclidean}\left(x,y\right)=\sqrt{\sum_{k=1}^{n}\left(x_k-y_k\right)^2} \tag{4}$$

Where x_k and y_k are the value of x and y on dimension n .

2.2.2. Manhattan Distance

Manhattan distance or City Block distance is defined as an addition of each its attributes. Hence, for two points of data x and y on dimension n, the distance between them is define as [19]:

$$d_{manha\tan}\left(x,y\right)=\sum_{k=1}^{n}\left|x_k-y_k\right| \tag{5}$$

Where x_k and y_k are the value of x and y on dimension n .

2.2.3. Chebisev Distance

Chebisev distance is also known as maximum distance. This distance is defined as the maximum value of distance from its attributes that exist. The distance for two points of data x and y on dimension n is define as [19]:

$$d_{chebisev}(x,y)=\max_{k=1}^{n}\left|x_k-y_k\right| \tag{6}$$

Where x_k and y_k are the value of x and y on dimension n .

2.2.4. Minkowski Distance

Minkowski distance is a measurement from metric distance which define as [24]:

$$d_{\min kowski}\left(p,x,y\right)=\sqrt[p]{\sum_{k=1}^{n}\left|x_k-y_k\right|^p} \tag{7}$$

With p is a Minkowski parameter. On Euclidean distance $(p=2)$, Manhattan distance $(p=1)$, and Chebisev distance $(p \to \infty)$. The metric condition for this function will satisfied as long as $p \geq 1$.

2.2.5. *Minkowski-Chebisev Distance*

This distance function is invented by Rodrigus (2018) and Novianty (2019) uses the combination of Minkowski-Chebisev distance for Fuzzy C-Means Clustering. The combination of Minkowski and Chebisev distance with weight w_1 and w_2 are shown on equation (7). When w_1 is greater than w_2, this distance function will looks similar to Manhattan distance and vice versa. The Minkowski-Chebisev distance is define as [25]:

$$\begin{aligned} & d_{\min-cheb}(w_1, w_2, p, x, y) \\ & = w_1 d_{\min kowski}(p, x, y) \\ & + w_2 d_{chebisev}(x, y) \end{aligned} \tag{8}$$

2.2.6. *Canberra Distance*

Canberra distance is an addition of the absolute ratio difference between two points of data [26]. The equation to measure this distance function is [27]:

$$d_{canbera}(x, y) = \sum_{k=1}^{n} \frac{|x_k - y_k|}{|x_k| + |y_k|} \tag{9}$$

This distance is very sensitive to the alteration if the two points of data are close to 0. This distance is applied because of its properties similarities with Manhattan distance.

2.2.7. *Average Distance*

Average distance is a modification of Euclidean distance. This modification is developed to improve the quality of clustering output [24]. This distance is define with this following equation [19]:

$$d_{average}(x, y) = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (x_k - y_k)^2} \tag{10}$$

2.3. *Cluster Evaluation*

Cluster Evaluation has to be applied to find out the accuracy level of the clustering algorithm and to perceive the existing classification. This research use 2 testing methods, Purity test and Accuracy test, for cluster evaluation [15].

3. Result and Discussion

Table 1 illustrate the detail information of the applied dataset.

Table 2. Dataset

No	*Dataset*	*Row*	Column	Class
1	Random 1	150	4	3
2	Random 2	210	7	3

3	Hill-Valley	1484	8	10
4	Iris	208	60	2
5	Seeds	606	100	2
6	Sonar	100	2	2
7	Yeast	400	2	4

Parameter values which used on this research are :$p = 5, w_1 = 0.5, w_2 = 0.5, m = 2, Max_{Iter} = 100, \varepsilon = 10^{-16}$. The values of objective function is observed in every iteration process. The output comparison of the clustering process on each dataset is done by normalizing the output value of the main objective by using this following equation (13):

$$x_i = \frac{x_i}{\max(X)} \times 100\% \quad (13)$$

Fig18 5 illustrate the result of the cluster evaluation average output in each of the distance which applied on Fuzzy C-Means Clustering algorithm. The result shows that the clustering process output on each distance function is not significantly different. This result is consistent with [6]. Nevertheless, the Euclidean distance and Average distance are achieving a better result than the other distance function. The same result is also achieved on another research [10].

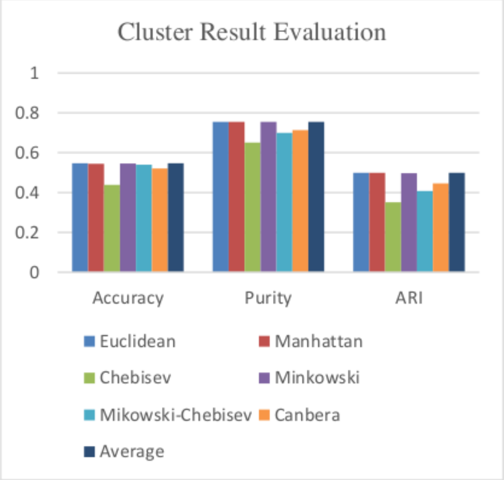

Figure 5. Cluster result evaluation

Apart from the desired clustering output, the low value of objective function which should be minimum is also the ideal result that we want to achieve. Figure 6 shows the average value of objective function which achieved on each distance function for 100 times iteration.

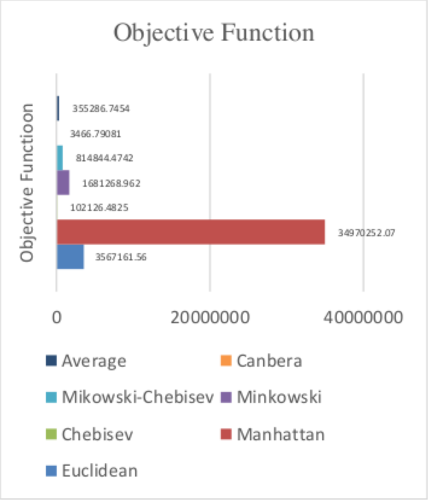

Figure 6. Objective function

The minimum values of objective function is achieved on 3 distance functions; Euclidean distance, Manhattan distance, and Minkowski-Chebisev distance. This result concludes that the combination of Minkowski distance and Chebisev distance able to improve the output of Fuzzy C-Means Clustering algorithm. This result also mentioned on different research [15] with 0.546 as the minimum value. The value of objective function on Manhattan distance is very different than other distance function which also shown in another research [9].

The other distance function parameter, the combination of Minkowski and Chebisev distance, does not produce the desired output, where the tendency of the objective function values are increasing on Yeast dataset, Sonar dataset, Seeds dataset, and Random 2 dataset. The most undesired result occurs in the application of Chebisev distance where the tendency of the objective function values are fluctuated. However, in different research, the usage of this distance function resulting on suitable performance when another parameters are added [13], [27]. In fact, the performance from the application of Chebisev distance still could be improved even further [9]. The results in this research are very dependent in the starting initialization value of the matrix. That's why the clustering method is also known as non-deterministic method. Nevertheless, the output of this research is considered beneficial for further analysis on how the Fuzzy C-Means Clustering method should be applied. Focusing on the clustering output is still not sufficient, the application of basic assumption where the values of objective function are always decrease also should be considered. For further research, it's essential to determine on how the termination criteria should be based on. The termination criteria in this research is based on the concept where the significant value of clustering result is constant (below the values of ε). This criteria is not exactly accurate, in fact, even when the minimum value of objective function is achieved, the objective function valued below ε are always continuously run the algorithm. On top of that, the possibility where the value of the objective function in the next iteration process got worse could be occurred.

4. Conclusion

This research concluded that the most suitable distance functions to apply for Fuzzy C-Means clustering method are Euclidean distance, Average distance, Manhattan distance, Minkowski distance, Minkowski-Chebisev distance, and Canberra distance. Furthermore, this research shows that not every distance function will satisfy the basic assumption where the values of objective function should be achieved even though the majority of the dataset that being applied meet the requirement. Based on the result of this research, it's important to visualize the tendency of the objective function values on Fuzzy C-Means clustering method. On the other hand, focusing on the result of the clustering process is also important. It aims to achieving suitable result through proper method. In the future, the research on the evaluation of the termination criteria and another different application of distance functions could be analyzed further.

References

[1] M. Zhang, W. Zhang, H. Sicotte, and P. Yang, "A New Validity Measure for a Correlation-Based Fuzzy C-Means Clustering Algorithm," pp. 3865–3868, 2009.

[2] N. Grover, "A study of various fuzzy clustering algorithms," *Int. J. Eng. Res.*, 2014.

[3] J. Nayak, B. Naik, D. P. Kanungo, and H. S. Behera, "ENGINEERING PHYSICS AND MATHEMATICS A hybrid elicit teaching learning based optimization with fuzzy c-means (ETLBO-FCM) algorithm for data clustering," *AIN SHAMS Eng. J.*, 2016, doi: 10.1016/j.asej.2016.01.010.

[4] J. Hernadi, *Analisis Real Elementer dengan Ilustrasi Grafis & Elementer*. Erlangga, 2015.

[5] G. R. Bartle and R. Sherbet, *Introduction to Real Analysis (4thed.)*. United States of America: Hamilton Printing Company, 2010.

[6] P. Grabusts, "The choice of metrics for clustering algorithms," *Proc. 8th Int. Sci. Pract. Conf.*, 2011.

[7] V. Kumar, J. K. Chhabrea, and D. Kumar, "Performance evaluation of distance metrics in the clustering algorithms," *INFOCOMP J. Comput. Sci.*, 2014.

[8] Y. S. Thakare and S. B. Bagal, "Performance evaluation of K-means clustering algorithm with various distance metrics," *Int. J. Comput. Appl.*, 2015.

[9] O. A. M. Jafar and R. Sivakumar, "Hybrid Fuzzy Data Clustering Algorithm Using Different Distance Metrics : A Comparative Study," no. 6, pp. 241–248, 2014.

[10] A. Singh, A. Rana, and U. Pradesh, "K-means with Three different Distance Metrics," vol. 67, no. 10, pp. 13–17, 2013.

[11] B. Charulatha, P. Rodrigues, and T. Chitralekha, "A comparative study of different distance metrics that can be used in Fuzzy Clustering Algorithms," *IJETTCS. Natl. Conf. Archit. Softw. Syst. Green Comput.*, 2013.

[12] V. P. Mahatme and K. K. Bhoyar, "Impact Of Distance Metrics On The Performance Of K-Means And Fuzzy C-Means Clustering-An Approach To Assess Student's Performance In E-Learning Environment," *Int. J. Adv. Res. Comput. Sci.*, 2018.

[13] S. Boddana and H. Talla, "Performance Examination of Hard Clustering Algorithm with Distance Metrics," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 2S3, pp. 172–178, 2019, doi: 10.35940/ijitee.b1045.1292s319.

[14] N. Gueorguieva, I. Valova, and G. Georgiev, "ScienceDirect ScienceDirect M & MFCM : Fuzzy C-means Clustering with Mahalanobis and Minkowski Distance Metrics," *Procedia Comput. Sci.*, vol. 114, pp. 224–233, 2017, doi: 10.1016/j.procs.2017.09.064.

[15] P. Noviyanti, "Fuzzy c-Means Combination of Minkowski and Chebyshev Based for Categorical Data Clustering," *Univ. Ahmad Dahlan*, 2018.

[16] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, "A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data," pp. 1–20, 2015, doi: 10.1371/journal.pone.0144059.

[17] J. Eliyanto, Sugiyarto, Suparman, I. Djakaria, A. H. Mustafa, and Ruhama, "Dimension Reduction Using Core and Reduct to Improve Fuzzy C-Means Clustering Performance," *Technol. Reports Kansai Univ.*, 2020.

[18] S. Surono and R. D. A. Putri, "Optimization of fuzzy c-means clustering algorithm with combination of minkowski and chebyshev distance using principal component analysis," *Int. J. Fuzzy Syst.*, 2020, doi: 10.1007/s40815-020-00997-5.

[19] G. Gan, C. Ma, and J. Wu, “Data clustering: theory, algorithms, and applications,” *Soc. Ind. Appl. Math.*, 2020.

[20] J. Han and M. Kamber, “Data Mining : Concepts and Techniques Third Edition,” *Morgan Kaufmann. Amsterdam*, 2011.

[21] J. C. Dunn, “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters,” *J. Cybern.*, 1973.

[22] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press, 1981.

[23] A. K. T. Andu and A. S. Thanamani, *Multidimensional Clustering Methods of Data Mining for Industrial Applications*. 2013.

[24] T. Brunello, D. Bianchi, and E. Enrico, *Introduction To Computational Neurobiology and Clustering*. Singapore: World Scientific Publishing, 2007.

[25] O. Rodrigues, “Combining minkowski and cheyshev: new distance proposal and survey of distance metrics using k-nearest neighbours classifier,” *Pattern Recognit. Lett.*, vol. 110, pp. 66–71, 2018, doi: 10.1016/j.patrec.2018.03.021.

[26] B. G. N. Lance and W. T. Williams, “Computer programs for hierarchical polythetic classification (" similarity analyses "),” 1964.

[27] C. Science, “Performance Evaluation of Distance Metrics in the Clustering Algorithms,” no. 1, pp. 38–51, 2014.

cek_Optimization Metric Space on Fuzzy C-Means Clustering

ORIGINALITY REPORT

14%	8%	13%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universiti Putra Malaysia Student Paper	2%
2	Submitted to University of Malaya Student Paper	1%
3	Sugiyarto Surono, Rizki Desia Arindra Putri. "Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis", International Journal of Fuzzy Systems, 2020 Publication	1%
4	Submitted to VIT University Student Paper	1%
5	Submitted to University of Southampton Student Paper	1%
6	es.scribd.com Internet Source	1%
7	Submitted to The Hong Kong Polytechnic University Student Paper	1%

8	Submitted to Wright State University Student Paper	1 %
9	"Harmony Search and Nature Inspired Optimization Algorithms", Springer Science and Business Media LLC, 2019 Publication	1 %
10	rhimrj.com Internet Source	<1 %
11	www.emerald.com Internet Source	<1 %
12	M. Y. Mashor. "Adaptive fuzzy c-means clustering algorithm for a radial basis function network", International Journal of Systems Science, 2001 Publication	<1 %
13	doaj.org Internet Source	<1 %
14	eprints.uad.ac.id Internet Source	<1 %
15	link.springer.com Internet Source	<1 %
16	www.docme.ru Internet Source	<1 %
17	www.ijert.org Internet Source	<1 %

18 CHAO-HSIEN CHU, JACK C. HAYYA. "A fuzzy clustering approach to manufacturing cell formation", International Journal of Production Research, 1991 <1%

Publication

19 Kaijie Xu, Witold Pedrycz, Zhiwu Li, Weike Nie. "Optimizing the prototypes with a novel data weighting algorithm for enhancing the classification performance of fuzzy clustering", Fuzzy Sets and Systems, 2021 <1%

Publication

20 M.N. Ahmed, S.M. Yamany, N. Mohamed, A.A. Farag, T. Moriarty. "A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data", IEEE Transactions on Medical Imaging, 2002 <1%

Publication

21 Manoranjan Paul, Manzur Murshed. "Impact of Clustering Techniques on Content-Based Pattern Generation Algorithm for Low Bit Rate Video Coding", 6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007), 2007 <1%

Publication

22 V. VENUGOPAL. "Soft-computing-based approaches to the group technology problem: a state-of-the-art review", International Journal of Production Research, 9/20/1999 <1%

Publication

23	dione.lib.unipi.gr Internet Source	<1%
24	docplayer.net Internet Source	<1%
25	hdl.handle.net Internet Source	<1%
26	journals.plos.org Internet Source	<1%
27	mafiadoc.com Internet Source	<1%
28	rdrr.io Internet Source	<1%
29	Amina Dik, Abdelaziz El moujahid, Abdelaziz Bouroumi, Aziz Ettouhami. "Weighted distances for fuzzy clustering", Applied Mathematical Sciences, 2014 Publication	<1%
30	R Khairi, S G Fitri, Z Rustam, J Pandelaki. "Fuzzy C-Means Clustering with Minkowski and Euclidean Distance for Cerebral Infarction Classification", Journal of Physics: Conference Series, 2021 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		