Analysis clustering using Normalized Cross Correlation in Fuzzy C-Means clustering algorithm

Ricky Crist Geoversam Imantara Kembaren[1)*], **Opim Salim Sitompul**[2)], **Sawaluddin**[3)]
[1,2,3)]Universitas Sumatera Utara, Medan, Indonesia
[1)]rickysembiring53@gmail.com, [2)]opiml@usu.ac.id, [3)]sawal@usu.ac.id

Submitted :Aug 9, 2022 | **Accepted** : Aug 27, 2022 | **Published** : Oct 3, 2022

Abstract: Fuzzy C-Means Clustering (FCM) has been widely known as a technique for performing data clustering, such as image segmentation. This study will conduct a trial using the Normalized Cross Correlation method on the Fuzzy C-Means Clustering algorithm in determining the value of the initial fuzzy pseudo-partition matrix which was previously carried out by a random process. Clustering technique is a process of grouping data which is included in unsupervised learning. Data mining generally has two techniques in performing clustering, namely: hierarchical clustering and partitional clustering. The FCM algorithm has a working principle in grouping data by adding up the level of similarity between pairs of data groups. The method applied to measure the similarity of the data based on the correlation value is the Normalized Cross Correlation (NCC). The methodology in this research is the steps taken to measure clustering performance by adding the Normalized Cross Correlation (NCC) method in determining the initial fuzzy pseudo-partition matrix in the Fuzzy C-Means Clustering (FCM) algorithm. the results of data clustering using the Normalized Cross Correlation (NCC) method on the Fuzzy C-Means Clustering (FCM) algorithm gave better results than the ordinary Fuzzy C-Means Clustering (FCM) algorithm. The increase that occurs in the proposed method is 4.27% for the Accuracy, 4.73% for the rand index and 8.26% for the F-measure..

Keywords: FCM, NCC, Clustering, Algorithm, Accuracy

INTRODUCTION

Fuzzy C-Means Clustering (FCM) has been widely recognized as a technique for data clustering, such as image segmentation (Pang et al, 2012). This technique is often also used for Data Mining needs in solving Big Data analysis (Xianfeng & Pengfei, 2015). There are several studies that have made improvements to solving problems in Fuzzy C-Means Clustering. One of these studies was carried out by (Khoshkbarchi et al, 2016), by combining the Fuzzy C-Means Clustering method with PSO or Particle Swarm Optimization. This study obtained an accuracy of 84.37% in testing the data. Meanwhile, FCM without combination obtained a lower accuracy of 80.58%. Another study was conducted by (Li, 2019), who conducted research on the modification of the FCM algorithm using Cosine Similarity. The modified algorithm in this case has a smaller error value and has a higher accuracy with an increase of 20.67% than the usual FCM algorithm.

This study will make improvements to the FCM algorithm using the Normalized Cross Correlation (NCC) method. In general, the FCM algorithm in determining the value of the initial pseudo-partition fuzzy matrix is carried out by a random process. The problem of using this method will be solved by the Normalized Cross Correlation (NCC) method.

*name of corresponding author

The Fuzzy C-Means Clustering algorithm still has weaknesses, such as: high sensitivity to initial conditions, no guarantee to achieve a global optimal solution and slow convergence to separate clusters. Thus, an additional method or improvement to the algorithm is needed to improve data clustering performance.

The purpose of this study was to analyze the performance of the Normalized Cross Correlation method on the Fuzzy C-Means Clustering algorithm in order to obtain better data clustering results. The test resulting from this method will be compared with the usual Fuzzy C-Means Clustering algorithm.

This study will conduct a trial using the Normalized Cross Correlation method on the Fuzzy C-Means Clustering algorithm in determining the value of the initial fuzzy pseudo-partition matrix which was previously carried out by a random process. The test results using the proposed method will be compared with the Fuzzy C-Means Clustering method without improvement and see the difference. This research will be tested on different datasets so that the performance of the method can be seen from several dataset sizes. The dataset test results will be analyzed and concluded based on the proposed improvements by finding the average value of the test results parameters from each dataset test obtained.

LITERATURE REVIEW

Clustering

Clustering is a common technique that is often applied in Data Mining, where this technique is used to group data into several clusters or groups based on the level of similarity between the data (Jain et al, 1999). Clustering technique is a process of grouping data which is included in unsupervised learning (unsupervised training). The process of grouping data like this has a simple nature and is similar to the human way of thinking. Humans do grouping of objects or data based on similarities or similarities to be grouped into the same group.

The right clustering results will obtain a high level of similarity in one class and a low level of similarity with other classes (Rossignol et al, 2018). The similarity in question is a numerical measurement of two objects or data. The similarity value between the two data will be higher, if the two data being compared have high similarity measurement results as well.

Clustering Methods

Data mining generally has two techniques in performing clustering, namely: hierarchical clustering and partitional clustering (Tan et al, 2006). The explanation of the method in data clustering

Hierarchical Clustering

This method is a technique of grouping data using a chart in the form of a hierarchy (levels). The grouping process is carried out based on combining data that has the closest resemblance in each iteration until all data sets are included in a cluster. Methods included in hierarchical clustering, namely: Single Linkage, Average Linkage, Complete Linkage, and Average Group Linkage.

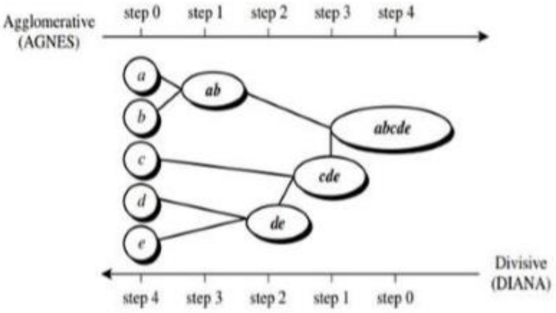

Fig 1 Hiearchical Clustering

In Fig 1 it can be seen how a data and other data will be grouped using a hierarchical chart based on the similarity of the data. Hierarchy-based clustering techniques can be divided into 2 types, namely: agglomerative nesting (AGNES) and divisive analysis (DIANA). Agglomerative clustering applies the process of grouping data using a bottom-up manner. This process begins by making each data from the

*name of corresponding author

dataset into a small cluster that has only one member. The next process, two clusters that have similarities will be grouped into one larger cluster.

Partitional Clustering Partitional

Clustering method is a technique of grouping data into several clusters without using a hierarchical chart such as the hierarchical clustering method. The partitional clustering method has a data center point for each cluster and in general has a goal of minimizing the distance from all data to each cluster center.

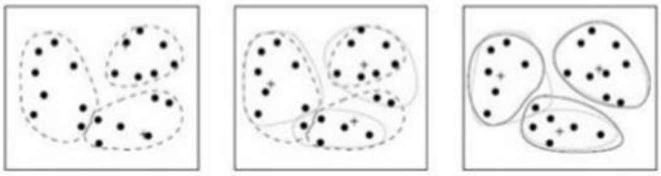

Fig 2 Partitional Clustering

In Fig. 2 the clustering process is carried out by determining a number of data that will be used as the initial cluster center point and finding or improving the cluster point through an iterative process. The results of these improvements can be in the form of final cluster points used in determining data clusters. Examples of methods in partitional clustering, namely: Fuzzy C-Means, K-Means and Mixture Modellin

Fuzzy C-Means Clustering (FCM)

Clustering technique is done by using a mathematical computational process, where the goal is to find structures or patterns related to similarities in one or more data and put them into data groups. A data can be grouped into one cluster by looking at the level of similarity that is identical. One algorithm that is quite popular in grouping the data is the Fuzzy C-Means Clustering (FCM) algorithm which is included in the partitional clustering method. The structure of the FCM algorithm has the possibility that each data attribute can be owned by more than one cluster with different degrees of membership. The degree of membership is a given parameter with values ranging from 0 (zero) to 1 (one).

In image data, FCM technique will group data or image pixels using fuzzy membership values. When the fuzzy membership value is 1 (one), it can be ascertained that the pixel intensity in that condition has similarities with the centroid value (data center) in the cluster (Xu et al, 2020). The performance of FCM is strongly influenced by the value of its membership which depends on the suitability of the approach between the data and the center of the cluster. This process causes the membership value and cluster center point to change (Tripathy et al, 2014).

Normalized Cross Correlation (NCC)

One of the techniques or methods applied to measure the similarity of data based on the correlation value is the Normalized Cross Correlation (NCC). The measurement results using the NCC method have an interval of 0 (zero) to 1 (one), where 1 (one) indicates the best match (Nakhmani & Tannenbaum, 2013). The use of this method can be used for various areas of need, such as: pattern recognition, template matching and signal processing.

METHOD

The methodology in this study is a step-by-step method for measuring clustering performance by adding the Normalized Cross Correlation (NCC) method in determining the initial fuzzy pseudo-partition matrix in the Fuzzy C-Means Clustering (FCM) algorithm. The clustering results obtained from the proposed method will be compared with the results of clustering using the usual Fuzzy C-Means Clustering algorithm by looking at the accuracy, rand index and F-measure values obtained.

The next process is to run the Fuzzy C-Means Clustering (FCM) algorithm and calculate the results of data clustering. The stages in this research can be described in Figure 3 below:

*name of corresponding author

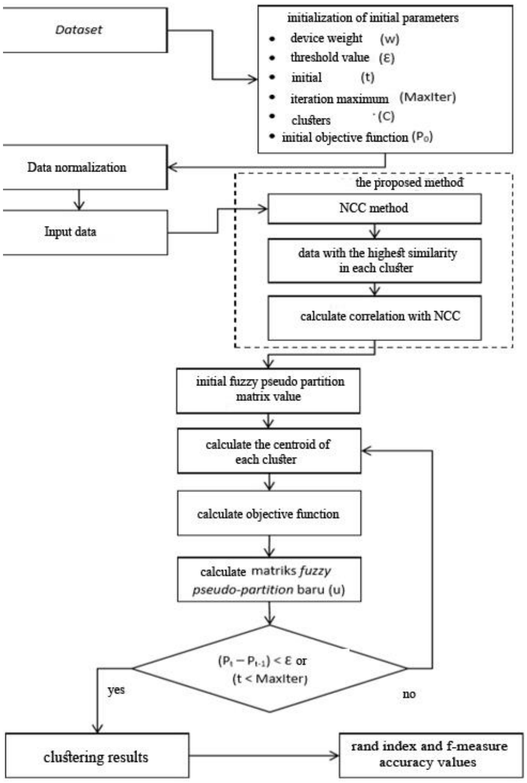
Dataset
initialization of initial parameters
device weight (w)
threshold value (Ɛ)
initial (t)
iteration maximum (MaxIter)
clusters (C)
initial objective function (P₀)
Data normalization
Input data
the proposed method
NCC method
data with the highest similarity in each cluster
calculate correlation with NCC
initial fuzzy pseudo partition matrix value
calculate the centroid of each cluster
calculate objective function
calculate matriks fuzzy pseudo-partition baru (u)
(Pₜ – Pₜ₋₁) < Ɛ or (t < MaxIter)
yes
no
clustering results
rand index and f-measure accuracy values

Fig 3 Block Diagram of *Clustering* with FCM and NCC Algorithms

The explanation of each stage carried out in the proposed research according to Fig. 3 can be described as follows, namely:

1) Conduct the selection process and enter the dataset that will be used as input data after normalization is carried out. The dataset will be divided into 2, namely: data attributes and data classes.
2) Determine the initial parameters of the FCM algorithm, namely:
    a) Weight of power (w = 2)
    b) Threshold value (Ɛ = 10-2)
    c) Initial iteration (t = 1)
    d) Maximum iteration (MaxIter = 100)
    e) Number of clusters (C = 2)
    f) The initial objective function (P0 = 0).
3) Perform the data normalization process using the min-max method. The data value for each attribute will be transformed into the same value with a range of values between 0 (zero) to 1 (one).
4) Finding the value of the initial pseudo-partition fuzzy matrix in each data group using the NCC method contained in equation (2.6). This value is obtained from the comparison of each data to find the highest correlation in each cluster using the NCC equation.
5) Calculate the value of the centroid between the data and the value of the initial fuzzy pseudopartition matrix.
6) Calculate the value of the objective function resulting from the input data, centroid and the value of the initial fuzzy pseudo-partition matrix.
7) Calculate the value of the new pseudo-partition fuzzy matrix (u) in each cluster. This process is carried out as an improvement or update of the previous pseudopartition fuzzy matrix value.
8) Check one of the following criteria, namely:
   If (Pt – Pt-1) < or (t < MaxIter), then the iteration stops
   If not, then t = t + 1 and repeat steps 4 to 7

9) Determine and calculate the results of data clustering obtained previously. These results state the cluster position of a data based on the maximum value obtained from the final pseudo-partition fuzzy matrix search results. The determination of the best clustering results can be measured using the confusion matrix technique. Values are sought and compared from the results of clustering obtained.

Data in this study amounted to 3 different datasets and were used to analyze the results of clustering on the Fuzzy C-Means algorithm which had been added with the Normalized Cross Correlation method. The dataset can be viewed from the internet page at the address https://archive.ics.uci.edu/ml/datasets.php. Brief information about the dataset used can be seen in Table 1 below.

Tabel 1 Datasets

No.	***Names of Data***	***Instances***	***Features***
1	Divoce Predictors	170	54
2	Breast Cancer Wisconsin	569	30
3	QSAR Biodegradation	1055	41

RESULT

Testing Results Dataset I

In testing dataset I, an initial initialization process will be carried out between FCM with the value of fuzzy pseudo-partition (u) which is determined randomly and FCM with value of fuzzy pseudo-partition (u) which is determined using the NCC method. The next step will be the clustering process on dataset I iteratively until it meets the termination criteria. Obtaining the initial objective function value in the formation of the fuzzy pseudo-partition (u) value using the ordinary FCM method, which is 786.13. Meanwhile, by using the FCM-NCC method, the initial objective function value is 636.60. Comparison of the acquisition of objective function values from the clustering process between the FCM and FCM-NCC methods in dataset I can be seen in Table 2 and Fig 4 below:

Tabel 2 Objective Results on Datasets I

Iteration to-	***Method***	
	FCM	**FCM_NCC**
Iteration 1	786.1305	636.5987
Iteration 2	675.3410	406.7657
Iteration 3	669.0547	311.4615
Iteration 4	618.8153	309.9059
Iteration 5	414.0368	309.8970
Iteration 6	311.7391	
Iteration 7	309.9065	
Iteration 8	309.8970	

*name of corresponding author

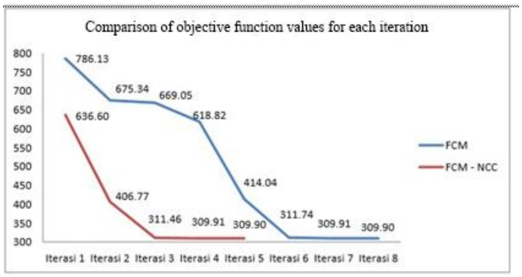

Fig 4 Comparison Graph of Objective Functions of Datasets I

The objective function in the first iteration for P1 will be compared with the value of P0 or the initial objective function of 0. In Tables 2 and Fig. 4 the clustering process ends in the 8th iteration for the ordinary FCM method with a difference in the value of the final objective function of 0.0095. Meanwhile, with the FCM-NCC method, the clustering process ends in the 5th iteration with the acquisition of the difference in the final objective function value of 0.0089. The results obtained in testing the dataset I showed that the results of clustering using the FCM-NCC method obtained better results with a smaller difference in the objective function.

By comparing the actual data label with the grouping results obtained, it can be calculated the value of accuracy, rand index and F-measure. These values will be compared to see which method is best used in clustering data. The results of the comparison of the methods used can be seen in Table 3 below:

Table 3 Results of Clustering with FCM and FCM-NCC on Dataset I

Information	***FCM***	***FCM_NCC***
Accuracy	97.65%	97.65%
Precision (P)	0.9765	0.9765
Recall (R)	0.9778	0.9778
Rand Index (RI)	0.9538	0.9538
F-Measure (F)	0.9771	0.9771

In Table 3, it can be seen that the comparison of the algorithms used produces the same clustering results. The FCM-NCC algorithm has the same accuracy, rand index and F-measure values as ordinary FCM.

Results of Testing Dataset II

In testing dataset II, the determination of the value of the initial pseudo-partition (u) fuzzy matrix is carried out like the process of testing dataset I. The clustering process on dataset II will be carried out iteratively until it meets the termination criteria.

Table 4 Results of Objective Functions on Dataset II

Iteration to-	***Method***	
	FCM	**FCM_NCC**
Iteration 1	177.2489	177.1824
Iteration 2	177.2249	175.1906
Iteration 3	177.2172	171.4111
Iteration 4		164.8850
Iteration 5		159.5439
Iteration 6		157.6635
Iteration 7		157.2711
Iteration 8		157.1977
Iteration 9		157.1834
Iteration10		157.1805

*name of corresponding author

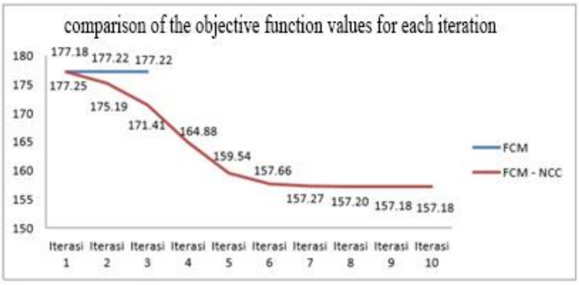

Fig 5 Comparison Graph of Objective Functions of Dataset II

In Table 4 and Fig 5, the clustering process will be carried out iteratively until it meets the termination criteria with the value of the difference in the final objective function being smaller than the value of (threshold).

The comparison between the actual data labels and the grouping results obtained is carried out to calculate the accuracy, rand index and F-measure values and see which method is best used in clustering data.

Table 5 Clustering results with FCM and FCM-NCC on Dataset II

Information	***FCM***	***FCM_NCC***
Accuracy	84.71%	92.79%
Precision (P)	0.8471	0.9279
Recall (R)	0.8353	0.9294
Rand Index (RI)	0.7405	0.8660
F-Measure (F)	0.8412	0.9287

NCC resulted in better clustering performance than without the NCC method. The FCM-NCC algorithm has better accuracy, rand index and F-measure values than ordinary FCM.

Results of Testing Dataset III

In testing dataset III, the process of determining the value of the initial pseudo-partition fuzzy matrix will be carried out as in the previous test. The clustering process will end for dataset III with the criteria for the difference in the value of the final objective function < (threshold) or iterations equal to the maximum value is met.

Table 6 Results of Objective Functions on Dataset III

Iteration to-	***Method***	
	FCM	**FCM_NCC**
Iteration 1	416.3249	359.9477
Iteration 2	359.7446	359.7108
Iteration 3	359.7127	359.7002
Iteration 4	359.7021	359.6963
Iteration 5	359.6979	

*name of corresponding author

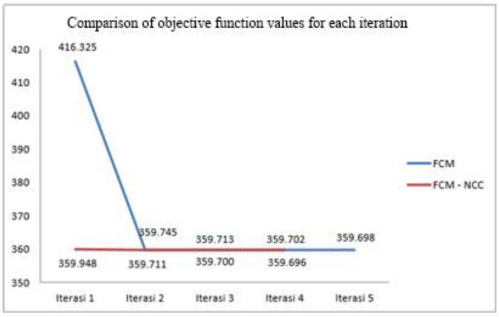

Fig 6 Comparison Graph of Objective Functions of Dataset III

In Table 4.5 and Figure 4.3, the clustering process ends after the difference in the value of the final objective function has a value that is smaller than the value of (threshold). The results of clustering with the usual FCM method ended in the 5th iteration with the acquisition of the difference in the final objective function value of 0.0042.

The last step in this experiment will be to compare the actual data labels with the grouping results obtained to determine the accuracy, rand index and F-measure values and see which method is best used in clustering data. The results of the comparison of the methods used can be seen in Table 6 below.

Table 6 Results of Clustering with FCM and FCM-NCC on Dataset III

Information	***FCM***	***FCM_NCC***
Accuracy	60.95%	65.69%
Precision (P)	0.6095	0.6569
Recall (R)	0.4726	0.7132
Rand Index (RI)	0.5324	0.5488
F-Measure (F)	0.85235	0.6839

In Table 6, it can be seen that the improvement of the FCM algorithm with the NCC method results in better clustering performance than without the NCC method..

DISCUSSIONS

Based on the results of clustering the data obtained, it can be seen and compared between the values of accuracy, rand index, and F-measure from testing the three datasets. The results of clustering will calculate the average value generated by adding up all test results divided by the total test. The average value obtained will be compared for testing the ordinary FCM algorithm with the FCM algorithm that has been added to the NCC method in determining the value of the initial pseudo-partition fuzzy matrix. The results of the data clustering test using the two methods can be seen in Table 7 and Table 8 below:

Table 7 Test Results with the FCM Algorithm

No.	***Datasets***	***Accuracy***	***Rand Index***	F-Measure
1	Datasets I	97.65%	0.9538	0.9771
2	Datasets II	84.71%	0.7405	0.8412
3	Datasets III	60.95%	0.5324	0.5235
Average result		**81.10%**	**0.7422**	**0.7806**

*name of corresponding author

Table 8 Test Results with the FCM-NCC Algorithm

No.	***Datasets***	***Accuracy***	***Rand Index***	F-Measure
1	Datasets I	97.65%	0.9538	0.9771
2	Datasets II	92.79%	0.8660	0.9287
3	Datasets III	65.69%	0.5488	0.6839
Average result		**85.38%**	**0.7422**	**0.8632**

In Table 7 and Table 8 above, the differences in clustering results can be seen from each test method used. The proposed method has a fairly good effect and results in the data clustering process. The average accuracy, rand index, and F-measure values obtained using the FCM algorithm are 81.10%, 0.7422 and 0.7806, respectively. While the average value of accuracy, rand index and F-measure obtained using the FCM-NCC algorithm, respectively, are: 85.38%, 0.7895 and 0.8632. The improvement in the proposed method is 4.27% for accuracy, 4.73% for rand index and 8.26% for F-measure.

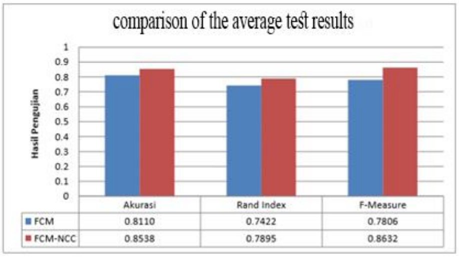

Fig 7 Comparison Graph of Clustering Data

In Fig 7, it can be seen that the results of accuracy, rand index, and F-measure for each dataset test using the FCM-NCC method get an increase in the results of data clustering. The next process, the data centroid value will be used for calculations on the value of the new pseudo-partition fuzzy matrix and the results will determine which class the data will be grouped into.

The next stage, each cluster center will be calculated its proximity to each data to obtain a similarity value and used as the initial value of the pseudo-partition fuzzy matrix. The clustering results obtained from the tests carried out state that the process of determining the value of the initial pseudo-partition fuzzy matrix in the FCM algorithm which generally uses a random or random method can also be carried out using the NCC method with better results.

1. CONCLUSION

Based on the testing and analysis, it can be concluded that the results of data clustering using the Normalized Cross Correlation (NCC) method on the Fuzzy C-Means Clustering (FCM) algorithm give better results than the usual Fuzzy C-Means Clustering (FCM) algorithm. The increase that occurs in the proposed method is 4.27% for accuracy, 4.73% for rand index and 8.26% for F-measure. The results obtained are that the determination of the initial pseudo-partition fuzzy matrix value in the FCM algorithm which generally uses the random or random method can also be done using the NCC method with better data clustering results.

REFERENCES

Fitriana, R. Saragih, J. & Luthfiana, N. 2017. Model business intelligence system design of quality products by using data mining in R Bakery Company. IOP Conference Series: Materials Science and Engineering.

Gueorguieva, N., Valova, I. & Georgiev, G. 2017. M&MFCM: Fuzzy C-means Clustering with Mahalanobis and Minkowski Distance Metrics. Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems - Procedia Computer Science.

Han, J., Kamber, M. & Pei, J. 2012. Data Mining: Concepts and Techniques.

Jain, A. K., Murty, M. N. &. Flynn, P. J. 1999. Data clustering: a review. ACM Comput. Surv. 31, 3 (Sept. 1999), 264–323.

Kaso, A. 2018. Computation of The Normalized Cross-Correlation by Fast Fourier Transform. PLoS ONE 13(9): e0203434.

Khoshkbarchi, A., Kamali, A., Amjadi, M. & Haeri, M. A. 2016. A Modified Hybrid Fuzzy Clustering Method for Big Data. 8th International Symposium on Telecommunications (IST).

Li, M. 2019. An improved FCM clustering algorithm based on cosine similarity. The 2019 International Conference - Association for Computing Machinery.

Nakhmani, A. & Tannenbaum, A. 2013. A New Distance Measure Based on Generalized Image Normalized Cross-Correlation for Robust Video Tracking and Image Recognition. Pattern Recognition Letters.

Pang, L., Xiao, K., Liang, A. & Guan, H. 2012. A Improved Clustering Analysis Method Based on Fuzzy C-Means Algorithm by Adding PSO Algorithm. Proceedings of the 7th international conference on Hybrid Artificial Intelligent Systems - Volume Part I.

Prasetyo, E. 2012. Data Mining Konsep dan Aplikasi Menggunakan Matlab. Yogyakarta : Andi Offset

Rossignol, M., Lagrange, M. & Cont, A. 2018. Efficient similarity-based data clustering by optimal object to cluster reallocation. PLoS ONE 13(6): e0197450.

Sano, A. V. D. & Nindito, H. 2016. Application of K-Means Algorithm for Cluster Analysis on Poverty of Provinces in Indonesia. ComTech: Computer, Mathematics and Engineering Applications.

Tan, P. N., Steinbach, M. & Kumar, V. 2006. Introduction to Data Mining. Boston : Pearson Education

Tripathy, B. K. 2014. Intuitionistic Fuzzy C means clustering with spatial information for image segmentation. ICCIC2014.

Uddin, A. M. 2014. Handwritten Bangla Character Recognition Using Normalized Cross Correlation. IOSR Journal of Computer Engineering. 16. 55-60.

Vijaya, Sharma, S. & Batra, N. 2019. Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 568-573.

Wen, Q., Yu, L., Wang, Y. & Wang, W. 2013. Improved FCM algorithm based on the initial clustering center selection. International Conference on Consumer Electronics, Communications and Networks.

Xianfeng, Y & Pengfei, L. 2015. Tailoring Fuzzy C-Means Clustering Algorithm for Big Data Using Random Sampling and Particle Swarm Optimization. International Journal of Database Theory and Application. 7. 191-202.

Xu, J., Zhao, T. & Feng, G. 2020. A Fuzzy C-Means Clustering Algorithm Based on Spatial Context Model for Image Segmentation. Int. J. Fuzzy Syst.