A Comparative Study Of Fuzzy C-Means And K-Means Clustering Techniques

Afirah Taufik
Department of Industrial Computing
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
firafify@gmail.com

Sharifah Sakinah Syed Ahmad
Department of Industrial Computing
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
sakinah@utem.edu.my

***Abstract*— Clustering analysis has been considered as a useful means for identifying patterns in dataset. The aim for this paper is to propose a comparison study between two well-known clustering algorithms namely fuzzy c-means (FCM) and k-means. First we present an overview of both methods with emphasis on the implementation of the algorithm. Then, we apply six datasets to measure the quality of clustering result based on the similarity measure used in the algorithm and its representation of clustering result. Next, we also optimize the fuzzification variable, m in FCM algorithm in order to improve the clustering performance. Finally we compare the performance of the experimental result for both methods.**

Keywords—clustering; fuzzy c-means; k-means; inter-class cluster

I. INTRODUCTION

Clustering analysis has been considered as useful means for identifying patterns of dataset [8]. Clustering is a technique for finding similarity groups in data, called clusters. The clustering algorithm will groups data into clusters based on the similarity among data. The goal of this process is to group the similar data in one group and the rest of data in other groups. Soft clustering is about assign a degree to which object in cluster and they are viewed as probability or score. Hard clustering is a partitioning the objects where the each objech in exactly one partition.

The fuzzy c-means clustering approach by Bezdek [1] are commonly used for fuzzy clustering. FCM allow each point to have degree with every cluster center. So that, each data points are given with a value between 0 and 1 membership to determine the degree of belonging to each group. The performance of FCM clustering depends on the selection of the initial cluster center and the initial membership value. The method are perform in iterative process. FCM using Sugeno training routine to make improvements in system modelling error [11].

K-means clustering algorithm is a hard clustering algorithm that can solve the well-known or well-separated clustering problem [2] [4]. K-means is about to find a k centroid for each cluster. The centroid should be place as much as possible from other centroid. These method is a simple clustering to classify the data to the nearest centroid [13]. It will keep updating until no longer data point can move. Determine the number of clusters in a data set, the quantity often labeled k, k-means algorithm, is a frequent problem in clustering data, and is an issue that is different from the process of solving problems of this group [10].

There are two similarity measure used for clustering performance, called intra-cluster and inter-cluster similarity. A good clustering method will produce clusters with a high intra-class similarity and low inter-class similarity [3]. The criterion of cluster validity is based on external and internal validity that represent best cluster for particular dataset. The most important of cluster validity is to test whether the data point are randomly being structured or otherwise. Beside the internal and external criteria the cluster validity also include the validity indices [5][6][7].

In FCM, many researchers use 2 as default for m, fuzziness coefficient or exponent for matrix U in FCM clustering. Alata et al. shows that value 2 is not optimal for all kind of research in many areas of clustering analysis. It must be depending on the dataset and the problems. Based FCM clustering algorithm, it needs the optimal number of clusters by optimize the parameters of the clustering algorithm by iteration search approach and then to find m, fuzziness coefficient, for the FCM algorithm. In order to get an optimal number of clusters, the iterative search approach is used to find the optimal single output Sugeno type fuzzy inference system (FIS) model to optimizing the algorithm parameters that give minimum error using the real data and Sugeno fuzzy model [9][13].

This paper aim to compare the clustering performance between hard clustering and soft clustering. Here the fuzzy clustering represent the soft clustering, whereas the k-means represent the hard clustering. In the rest of the paper we describe in detail the clustering techniques and the quality measure in Section II and give the experimental result in Section III.

II. TECHNIQUES AND QUALITY MEASURE

In this section, we discuss two clustering methods for evaluation the clustering results, known as fuzzy c-means and k-means.

A. Fuzzy c-means

- FCM is a clustering method which allows one piece of data to belong to two or more clusters. It is based on the minimization of the objective function can see in (1).

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

- m is the exponent for the partition matrix U and range is from 1 to 3. u_{ij} is a membership degree of x_i in the cluster j, x_i is the ith of measured data, c_j is the centre of the cluster, and $\| \ \|$ is any norm that measured the similarity between data in centre cluster. FCM are executed through an iterative optimization of the objective function algorithm shown above, and also update membership degree u_{ij} (2) and the point of cluster centres c_j by this function in (3).

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

$$c_j = \frac{\sum_{i-1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \qquad (3)$$

- This iteration will stop when (4).

$$\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon, \qquad (4)$$

B. K-means

The main idea is to define k centroids, one for each cluster. The next step is to take each point owned by the data given set and cluster it to the nearest centroid. Then compute a new centroid k barycentre clusters resulting from the previous step. After we have these k new centroids, a new bind must be done between similar dataset points and the closest new centroid. A updating and iteration has been generated. As a result of this process that the k centroids change their location step by step until there is no longer move. In other words, centroids are the last one. Finally, this algorithm aims to minimize the objective function, in this case the squared error functions. The objective function can see in (5).

- $J = \sum_{j=1}^{k} \cdot \sum_{i=1}^{n} \left\| x_i^{\ j} - c_j \right\|^2$ (5)

- Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centers.

The algorithm can best summarized by the following steps:

- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
- Assign each object to the group that has the closest centroid.
- When all objects have been assigned, recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

C. Inter-class cluster similarity

The inter-class and intra-class cluster similarity is a crucial part in clustering. The validity of good clustering can be test based on this experiment. In figure 1 show the difference between inter-class and intra-class. The inter-class cluster show the distance between data point with cluster center, meanwhile intra-class cluster show the distance between the data point of one cluster with the other data point in other cluster.

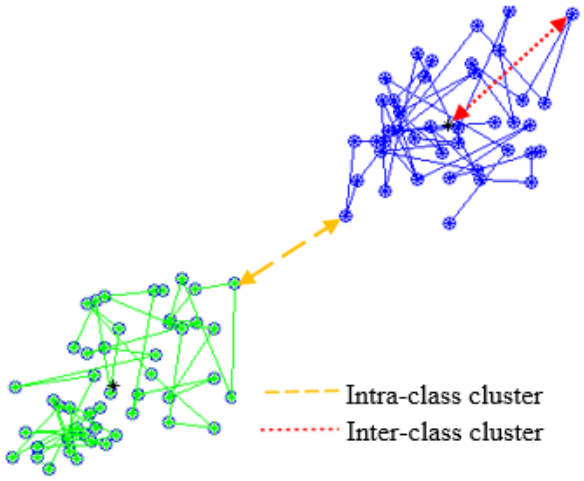


Figure 1. Example of inter-class and intra-class cluster similarity

In this paper, we emphasis only for inter-cluster similarity. This experiment is about to compare between these two methods FCM and k-means clustering by the distance of inter-class cluster similarity. The data are similar, and nearest to each other's belongs to one cluster. The results are based on total summation distance by summation all the distance within one clusters of the dataset. The comparison uses the same number of cluster for both two methods FCM and k-means to execute a function on the dataset. For FCM, the summation are getting from maximum U in every cluster and after that, divide with the total vectors. For k-means are calculate based on total summation of distance and divide with the total of vectors.

D. Reconstruction error

This experiment are based on FCM algorithm, and the optimal m is based on the minimum error. The error are called as reconstruction error. At (7) is the formula to construct an error and a new $\hat{x}$. When getting the new x based on (6) formula. Next get the error for every different m exponent. The m values change with different value range 1 to 3. Later, the experimental result for a reconstruction error is to find m optimal based on minimum error is executed in the next chapter using the ten generated synthetic dataset [12].

$$\hat{x} = \frac{\sum_{i=1}^{c}(\tilde{u_i})v_i}{\sum_{i=1}^{c}(\tilde{u_i})^m} \tag{6}$$

$$\text{Error} = x - \hat{x} \tag{7}$$

III. Results and Experimental

A. Synthetic Datasets

In table 1, four real datasets were obtained from UCI Machine Learning Repository and two synthetic datasets are listed, and the total of vectors are show. These dataset contains a characteristic such as a borderline data, overlapping data and wellseparate data to execute with the experiments.

Table 1: Six synthetic dataset in 2-d

Name of	Total of
S2	5000
S3	5000
Flame	240
Pathbased	300
Ds3	85
Ds4	300

B. Comparison of FCM and K-means

First experiment starts with s2 dataset and for both methods that are FCM and k-means executed in Matlab. As can see in table 2 for s2 data, s2 is a scattered dataset and have some data are close to each other at certain point. S2 data have 5000 vectors. The number of clusters chosen is 4. In figure 1(a) show a membership degree for each data point to the four center cluster. The highest membership degree is at the 4th row of the first column means the data point have a high membership with a second cluster. It was showed that for the first data in s2 dataset belonged to cluster number 4 clusters. It is the same way for the next 5000 vectors. Column is representing a data and a row with the highest degree is representing in which cluster they are. The membership degrees of FCM show the fuzzy value set. The equal summation of one column is 1. Here we can see that the data point are belong to all cluster based on the membership degree.

For k-means, the visualization in figure 1(b) also use the same dataset, s2, the scattered data in dataset was plotted based on k-means with the number of clusters is 4. Additional, the index of data show they are belonging to which cluster. The first row means the first data belongs to cluster 1. It is the same way for the next 5000 vectors. Comparison between these two clustering method is the membership degree of FCM and k-means are based on fuzzy or crisp value table as can see in figure 1. The division of the membership function of k-means is only 0 and 1, crisp set. If the data belong to that cluster, the value is 1 or otherwise.

0.110833	0.122358	0.13192	0.115561
0.326223	0.255689	0.224906	0.343065
0.093062	0.092521	0.09309	0.09822
0.469882	0.529432	0.550085	0.443154

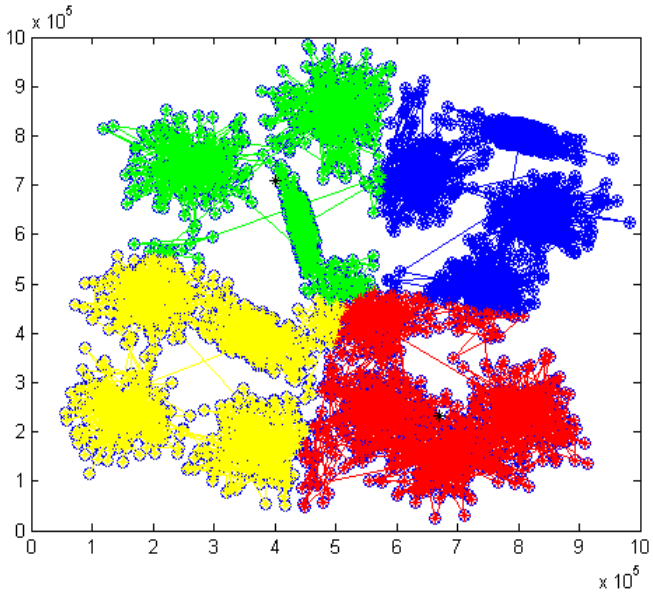
(a)

1	1	3	1

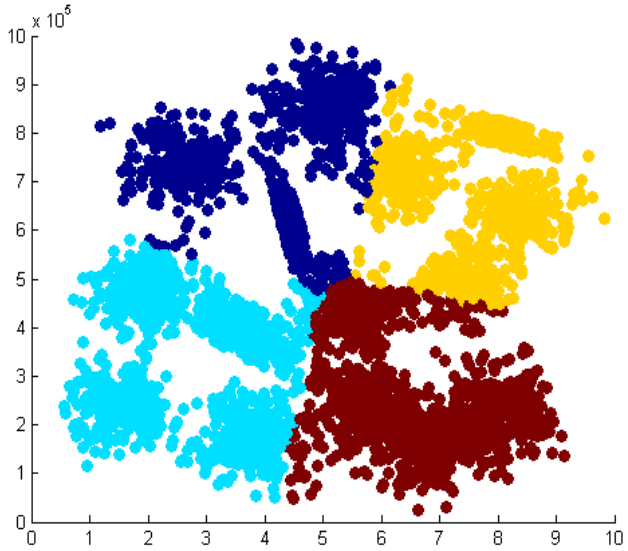
(b)

Figure 1. (a) Membership degree of s2 dataset (b) Index of s2 dataset for k-means

Figure 2(a) are show the visualization of s2 dataset in four cluster for FCM. In this figure, the data point are clustered based on the highest membership degree. The s2 data are complex and FCM helps to solve the data point at borderline using membership degrees. Figure 2(b) show the visualization for k-means in four clusters. For a s2 dataset, k-means are not suitable because the s2 is a complex data and every data point are need a knowledge such a membership degree given by FCM. From the graph we can see both algorithm give similar clustering result.



(a)



(b)

Figure 2. Synthetic data S2 with k = 4 (a) FCM (b) k-means

Figure 3 show the membership degree of experiment with flame dataset. This data contains 240 vectors. It was clustered to two clusters only. At figure 4 for dataset flame, the visualization shows the separation of two clusters also has a borderline case just at a certain point. At figure 3(a), the third column shows the more or less degree than the other one membership degree. The FCM helps to solve data at borderline case. Even though this dataset have a certain point of borderline cases, this data point is a scattered data that have a separation with each other. This data are also suitable used with k-means clustering algorithm because there is not a complex dataset and do not have many vectors. In figure 4, dataset showed the visualization for k-means and the index of which cluster can see in figure 3(b). It goes the same for other four dataset. The chosen number of cluster for those dataset - s3, pathbased, ds3 and ds4 is 6, 2, 2 and 2 respectively. The visualization of plotted graph is in table 2.

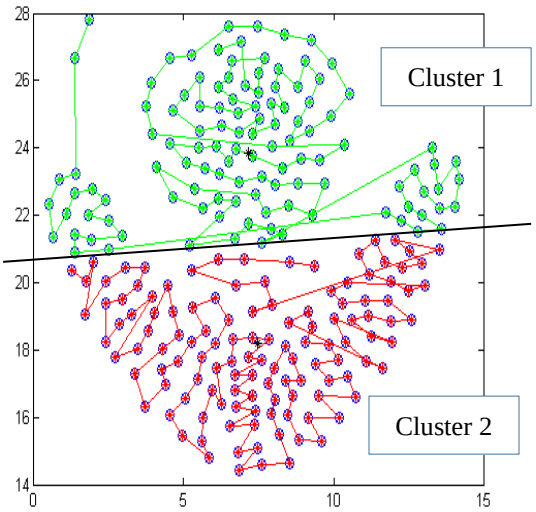
0.7000058	0.6378923	0.5944086	0.6130863
0.2999942	0.3621077	0.4055914	0.3869137

(a)

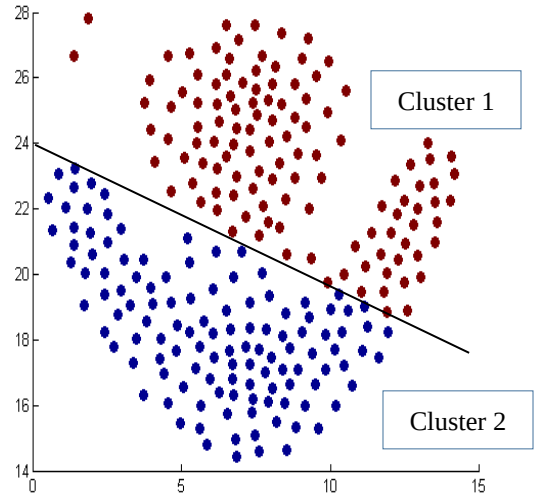
2	2	1	1

(b)

Figure 3. (a) Membership degree of Flame dataset (b) Index of Flame dataset for k-means



(a)



(b)

Figure 4. Flame data with k = 2 (a) FCM (b) k-means

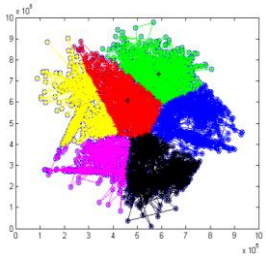
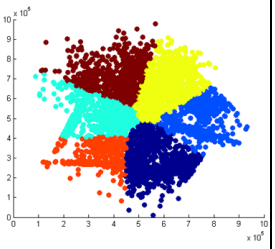
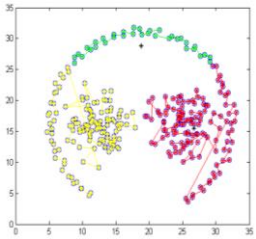
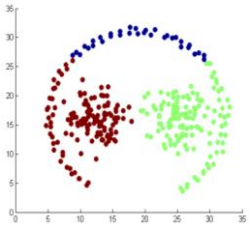
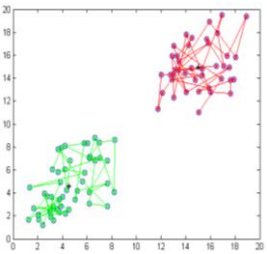
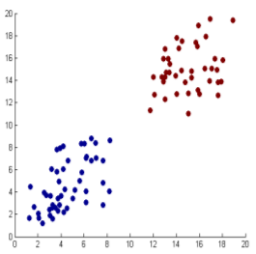
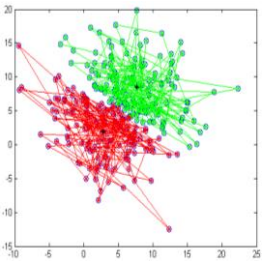
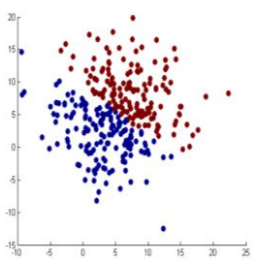
The summarization can be conclude based on this six experimental with different synthetic 2-d dataset and using the two methods algorithm of FCM and k-means. The s2 and s3 dataset have a 5000 vectors and this data show the complex data. By choose the number of clusters, the visualization pattern of dataset such as a borderline, outliers and overlapping cases can be seen. The knowledge of the membership function can show a very detail exactly which clusters that every data point belong to. By see the results of visualization for k-means and the index which cluster are the data point belongs to, just can see the simple of results and not details such a membership degree of FCM. As mention above in the literature review part, k-means are about to solve Apriori specification based on the centroid (cluster center).

Based on the results of experimental of flame, pathbased, and ds4 dataset show that in one dataset that have a scattered data also can have such as separated data and borderline case of data. These problems showed that the knowledge of the

membership degree is very important to classify the different pattern of data in clustering. The membership degree will define more or fewer grades or much-separated grades of the data. The FCM helps to present the nature of clustered data.

Based on the results of experimental of ds3 dataset, shows that these dataset have a very well separate data can see by the visualization of the results. This kind of dataset is to prove that it is suitable to execute with k-means clustering algorithm. This kind of dataset is not necessarily needs to execute with FCM, otherwise the knowledge of the membership function are needs or to execute with other dimensional data, so the membership degree of FCM are important.

Table 2: Comparison scattered data for FCM and k-means

Data set	FCM	K-means
S3		
Path-based		
DS3		
DS4		

C. *Summation distance for inter-cluster for FCM and K-means*

The experiments are used the same four real datasets and two synthetic datasets, s2, s3, ds3, ds4, flame and pathbased. The experiment was executed used both methods. The results as can see in Table 3.

In order to obtain the summation of inter-cluster center, functions in Matlab are executed to get the center of a cluster based on the number of clusters. After the process of updating and iteration until the place of center are no longer move (only for k-means), whereas for FCM the initial cluster center and membership are initialize. After that, the data point are clustered based on similarity of the data and to the nearest center cluster whereas. The function will sum all the distance of the data in one cluster. After that, the function will give a final score for the summation of distance for a particular dataset. Based on table 3, the results show that FCM for the summation distance are smaller than k-means summation distance for all dataset. The good score (low value of a summation of distance) show that in one cluster in one dataset have higher memberships between data point and the cluster center.

Table 3: Summation distance for inter-cluster for FCM and k-means

Dataset	FCM	K-means
S2	$\mathbf{2.3965 \times 10^5}$	2.3958×10^{10}
S3	$\mathbf{2.0096 \times 10^5}$	1.1276×10^{10}
Flame	**2.8783**	13.0157
Pathbased	**10.8455**	29.8597
Ds3	**6.8302**	8.1886
Ds4	**3.6798**	35.1584

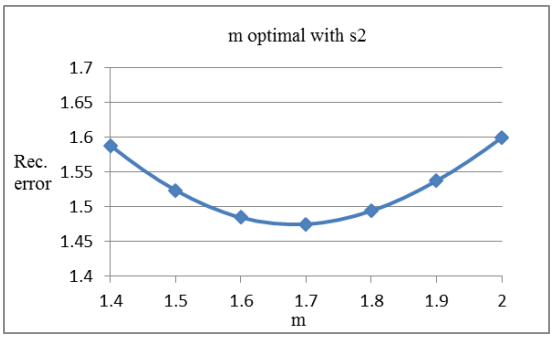
D. *Finding m optimal based on reconstruction error*

The experiment continues by the function of reconstruction error to find optimal m, exponent by using different value m for FCM. The range of is m, about 1 to 3. m is the exponent for the partition matrix U in FCM. In FCM clustering application, many researchers used 2 as a typical m. The optimal m obtain based on the minimum reconstruction error after executed with different value m, with much time of the run. The graph was plotted to show the optimal m for every dataset. The dataset used for this experiment is the synthetic 2-dimensional same as the previous experiment.

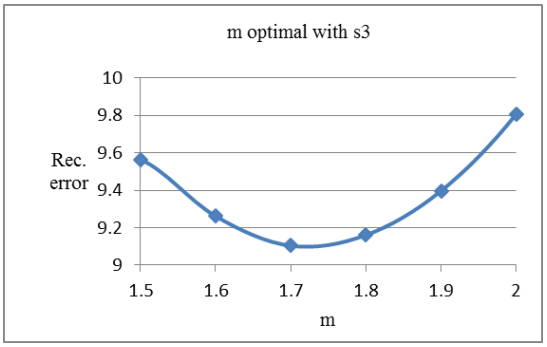
The reconstruction error means construct a new x from the previous x based on the center and the membership degree of data points to the center cluster. After getting the new x, will be deducted with the previous x and get an error. The error is record for various number of m. The optimal m are getting from minimum errors after many runs in Matlab.

The dataset s2 and s3 are a complex data. Based on the experiment for a reconstruction error, the optimal m exponent for these two dataset is 1.7. For a flame and pathbased dataset, these data have a pattern as can see in the visualization for the previous experiment. These two data are scattered data and be clustered to the number of clusters depends on the pattern of

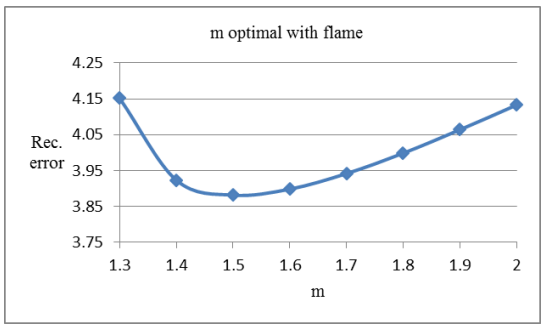
the data. The optimal m for these two dataset is 1.5 and 1.6. For a well-separated data, such ds3 is 2.1. For a ds4 dataset that have points at borderline and the separated data get the optimal m for this experiment is 1.5. The optimal m can be view by the results on figure 4.



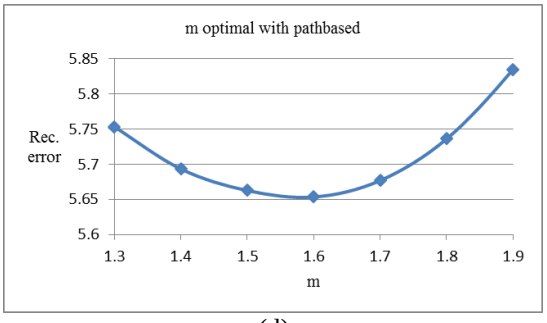
(a)



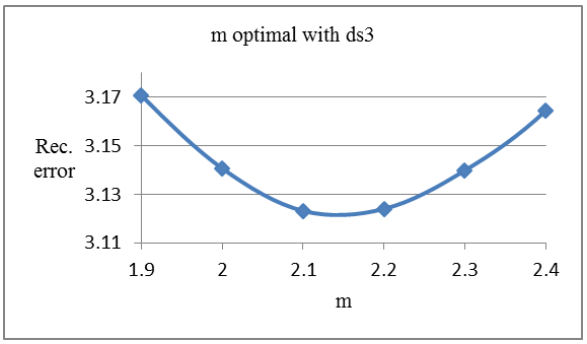
(b)



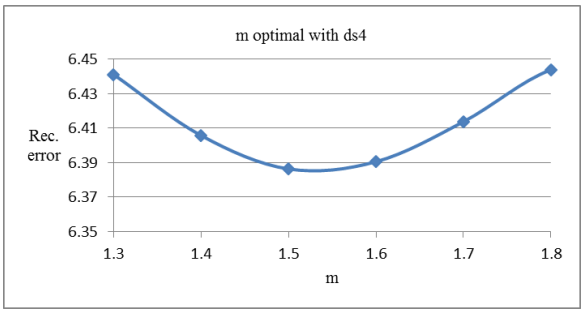
(c)



(d)



(e)



(f)

Figure 4. m optimal by minimal reconstruction error

IV. Conclusion

As a conclusion, for experimental to compare between both methods, the FCM is better than k-means that the membership degree of FCM can give knowledge and exactly information in fuzzy set, about the data either the data is complex or simple data. The understanding of row and column for membership function is also helping to know which data are belong to which cluster. The k-means is suitable to solve a simple and well-separated dataset which is can see by the visualization of plotted graph for k-means by the index give which cluster are the data belongs to.

The experiment for inter-class cluster similarity are based on low-value summation of distance. Based on the results, FCM have a good score of the inter-class cluster similarity. The good score (low value of a summation of distance) show that in one cluster in one dataset have higher memberships between data point and the cluster center.

The reconstruction error experiment can help to find the m optimal for every dataset and also to prove that the value m exponent is can be any number range 1 to 3 including the typical m, exponent = 2. This matter can prove that not all the research field of clustering suitable with default m. This experiment proves that to get these values of optimal m also depends to the dataset and the problems.

ACKNOWLEDGMENT

This research was supported by Universiti Teknikal Malaysia Melaka and Ministry of Education Malaysia under grant no FRGS/2013/FTMK/ICT02/02/1/F00161.

REFERENCES

[1] James C.Bezdek, R.E., William Full, "FCM: The Fuzzy c-Means Clustering Algorithm" 1984.

[2] J. MacQueen "Some Methodss for Classification and analysis of Multivariate Observations" 1967.

[3] Guha, S., Rastogi, R., and Shim K. "CURE: An Efficient Clustering Algorithm for Large Databases" 1998. In Proceedings of the ACM SIGMOD Conference.

[4] Kiri Wagsta, Claire Cardie, Seth Rogers, Stefan Schroedl "Constrained K-means Clustering with Background Knowledge" Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577-584.

[5] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis, "On Clustering Validation Techniques" Journal of Intelligent Information Systems, 17:2/3, 107–145, 2001.

[6] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis, "Cluster Validity Methods : Part 1" SIGMOD Record, vol.31, no.2, June 2002.

[7] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis, "Clustering Validity Checking Methods : Part II" SIGMOD Record, vol.31, no.3, Sept 2002.

[8] R. J. Almeida, J.M.C.S., "Comparison of fuzzy clustering algorithms for classification", 2006.

[9] Mohanad Alata, M.M., Abdullah Ramini, "Optimizing of Fuzzy C-Means Clustering Algorithm Using GA" 2008.

[10] Pakhira, M.K., "A Modified k-means Algorithm to Avoid Empty Clusters" 2009.

[11] K.M. Bataineh, M.Naji, M.Saqer "A Comparison Study between Various Fuzzy Clustering Algorithms" vol.5 no.4 Aug. 2011.

[12] W. Pedrycz, S.S. Syed Ahmad, Evolutionary feature selection via structure retention, Expert Systems with Applications, Volume 39, Issue 15, 1 November 2012, Pages 11801-11807

[13] S.S. Syed Ahmad, "Fuzzy Modeling through Granular Computing", University of Alberta, 2012