Comparing Fuzzy-C Means and K-Means Clustering Techniques: A Comprehensive Study

Sandeep Panda, Sanat Sahu, Pradeep Jena, and Subhagata Chattopadhyay

Dept. of Computer Science and Engineering
National Institute of Science and Technology
Palur Hills, Berhampur 761008 Odisha India
{sandeeppandakumar15081991,sanat.lipu,
subhagatachatterjee}@gmail.com, pradeep1_nist@yahoo.com

Abstract. Clustering techniques are unsupervised learning methods of grouping similar from dissimilar data types. Therefore, these are popular for various data mining and pattern recognition purposes. However, their performances are data dependent. Thus, choosing right clustering technique for a given dataset is a research challenge. In this paper, we have tested the performances of a Soft clustering (e.g., Fuzzy C means or FCM) and a Hard clustering technique (e.g., K-means or KM) on Iris (150 x 4); Wine (178 x 13) and Lens (24 x 4) datasets. Distance measure is the heart of any clustering algorithm to compute the similarity between any two data. Two distance measures such as Manhattan (MH) and Euclidean (ED) are used to note how these influence the overall clustering performance. The performance has been compared based on seven parameters: (i) sensitivity, (ii) specificity, (iii) precision, (iv) accuracy, (v) run time, (vi) average intra cluster distance (i.e. compactness of the clusters) and (vii) inter cluster distance (i.e. distinctiveness of the clusters). Based on the experimental results, the paper concludes that both KM and FCM have performed well. However, KM outperforms FCM in terms of speed. FCM-MH combination produces most compact clusters, while KM-ED yields most distinct clusters.

Keywords: Clustering, FCM, KM, Distance measures, Performance test.

1 Introduction

Clustering is a method of grouping similar data and distinctly separating them from the dissimilar data. It helps recognizing hidden patterns within the data. It is an unsupervised approach. For pattern extraction, clustering techniques depend on the similarity measures between the representative and the data to be clustered. Representative data denotes the cluster center, i.e., the ideal data of the cluster. Similarity is computed based on the distance measure between the cluster center and the data to be clustered using several methods, such as Manhattan, Euclidean, Cosine, Mahalanabis, and Hamming etc. The advantages of clustering techniques are that these do not require domain knowledge and labeled data, are able to deal with various types of data (including noisy data and outliers), capable of interpreting ad-hoc data and could be reused.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 451–460.
springerlink.com © Springer-Verlag Berlin Heidelberg 2012

There are two broad types of clustering methods, e.g., 'Soft' and 'Hard' clustering. Soft clusters are devoid of distinct boundaries, as seen in Fuzzy C Means (FCM) [1], Fuzzy K-nearest Neighbor (FKN) [2], Entropy-based Fuzzy Clustering (EFC) [3] and so on. On the other hand, 'hard' clusters possess well-defined boundary, which is seen in K-means (KM) [4], Hierarchical methods [5] and so forth. Choosing correct algorithm has always been a research challenge [6]. In this paper, we compare the performance of one 'soft' (e.g., FCM) and one 'hard' clustering i.e., KM technique on three standard datasets of various sizes. These datasets are *Iris* (150 x 4), *Wine* (178 x 13) and *Lens* (24 x 4), obtained from UCI machine learning database [7]. Performances of FCM and KM are also compared based on Manhattan (MH) and Euclidean (ED) measures.

The objective of this study is to examine the best *'clustering methods-distance measure'* combinations in terms of (a) 'quantitative clustering' (which are checked with Sensitivity, Specificity, Precision and Accuracy measures); (b) 'speed' (examined by measuring the run time); and (c) 'quality' of the cluster in terms of compactness and distinctiveness (i.e., how far one cluster is situated from another cluster).

2 Methodology

The objective of this study is to compare the performance of FCM and KM on three standard datasets, such as Iris, Wine and Lens. In order to accomplish the task, the algorithms are developed in 'C' language and implemented.

Working Principle of FCM Algorithm:

1. *Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$*
2. *At k-step: calculate the centroid $C^{(k)}=[c_j]$ with $U^{(k)}$*

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

3. *Update $U^{(k)}$, $U^{(k+1)}$*

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

4. *If $\| U^{(k+1)} - U^{(k)} \| < \theta$ then STOP; otherwise return to step 2.Here, 'm' is the fuzziness parameter.*

Working Principle of KM Algorithm:

For 'M' sample vectors $\{x_1,x_2,\ldots,x_M\}$ falling into 'k' compact clusters (k<M)

```
    Let '$m_i$' be the mean of the vectors in cluster 'i'
    If $\|x - m_i\|$ is the minimum of all 'k' distances
    insert 'x' into the respective cluster
            until there is no change in any 'm'.
```

To note how the distance measures influence the clustering tasks, two distance measures have been used, e.g., Manhattan (MH) and Euclidean (ED). These distances follow L^P -norm (see equation 1).

$$\|x\|_p = \left[\sum_{i=1}^{k} |x_i|^p\right]^{\frac{1}{p}} \tag{3}$$

In this equation, 'x_i' are the number of data points. Now for 'p'=1 we get MH and for 'p'=2 it is ED.

The performance of these two techniques has been compared based on the following parameters:

1. Sensitivity: $\frac{T_P}{P}$ (4)

In this equation T_P is 'true +' and 'P' is '+'.

2. Specificity: $\frac{T_N}{N}$ (5)

   In this equation T_N is 'true -' and 'N' is '-'.

3. Precision: $\frac{T_P}{T_P + F_P}$ (6)

   In this equation Fp is 'false +'

4. Accuracy: $\frac{n}{n*}$ (7)

   Here 'n' denotes the total number of correctly classified data and 'n*' is the total number of data.

5. Run time: the amount of time (in seconds) spent to run the algorithm in a PIV (core2duo) PC.

6. Intra cluster distance: $\frac{1}{n*}\sqrt{\sum_{i=1}^{n*}\|x_i - c\|^2}$ (8)

   In this equation, 'c' denotes the cluster center (or, centroid). From this equation, we can infer if such distance is low, the respective clusters are more 'compact'.

7. Inter cluster distance: $d = \left[\left\| c_{ij} - c_{ji} \right\|^{p} \right]^{\frac{1}{p}}$ (9) $(i \neq j); p = 2$

   Here, ‘m’ is the desired number of clusters. The desired inter cluster distance should be high to infer that the clusters are not overlapped.

3 Results and Discussions

In this section, the experimental results are displayed and explained as follows. Table 1-3 shows the performance results on three datasets.

Table 1. Performance analysis of FCM and KM clustering techniques on **Iris** data

		FCM		**KM**	
Parameters	**Cluster-info**	**MH**	**ED**	**MH**	**ED**
Sensitivity	CL1	1	1	1	1
	CL2	0.94	0.94	0.84	0.84
	CL3	0.72	0.72	0.84	0.84
	Average	0.8866	0.8866	**0.8933**	**0.8933**
Specificity	CL1	1	1	1	1
	CL2	0.86	0.86	0.82	0.82
	CL3	0.97	0.97	0.9	0.9
	Average	**0.9433**	**0.9433**	0.9066	0.9066
Precision	CL1	1	1	1	1
	CL2	0.7705	0.7705	0.84	0.84
	CL3	0.9231	0.9231	0.84	0.84
	Average	**0.8978**	**0.8978**	0.8933	0.8933
Accuracy	*Average*	0.8867	0.8867	**0.8933**	**0.8933**
Time	*Average*	0.2321	0.2321	**0.1281**	**0.1281**
Intra cluster Distance	CL1	0.895	0.487	1.261	**0.696**
	CL2	1.261	0.695	0.908	**0.493**
	CL3	1.184	0.65	1.101	**0.605**
Inter cluster Distance	CL1-CL2	**1.487**	0.775	1.47	0.767
	CL2-CL3	**2.129**	1.174	0.881	0.448
	CL1-CL3	**0.951**	0.487	2.012	1.107

Table 2. Performance analysis of FCM and KM clustering techniques on **Wine** data

		FCM		**KM**	
Parameters	**Cluster-info**	**MH**	**ED**	**MH**	**ED**
Sensitivity	CL1	1	1	1	1
	CL2	0.9014	0.9014	0.9155	0.9155
	CL3	1	1	1	1
	Average	0.9671	0.9671	**0.9718**	**0.9718**
Specificity	CL1	0.9748	0.9748	0.9748	0.9748
	CL2	1	1	1	1
	CL3	0.9692	0.9692	0.9769	0.9769
	Average	0.9813	0.9813	**0.9839**	**0.9839**
Precision	CL1	0.9516	0.9516	0.9516	0.9516
	CL2	1	1	1	1
	CL3	0.9231	0.9231	0.9412	0.9412
	Average	0.9582	0.9582	**0.9642**	**0.9642**
Accuracy	*Average*	0.9607	0.9607	**0.9663**	**0.9663**
Time	*Average*	0.5007	0.5007	**0.1976**	**0.1976**
Inter cluster Distance	CL1	**2.28**	0.772	2.2109	0.7430
	CL2	**3.523**	1.105	2.6007	0.8259
	CL3	**2.701**	0.85	3.4510	1.0867
Intra cluster Distance	CL1-CL2	**2.29**	0.786	2.5148	0.8338
	CL2-CL3	**2.564**	0.849	2.2870	0.7825
	CL1-CL3	**2.797**	0.849	2.7950	0.9166

In Iris data, KM clusters with greater accuracy than FCM, which renders negligibly better precision. While testing the run time, it may note that KM takes much less time compared to FCM for both the MH and ED distances. FCM-ED is able to produce most compact clusters, while FCM-MH yields most distinct clusters. In Wine data, again the first five parameters (sensitivity, specificity, precision, accuracy, and run time) show similar results as seen in Iris data. KM-ED combination is able to produce the most compact clusters. Both KM-MH and FCM-MH is able to produce most distinct clusters. Similar results (as seen in Wine) could be seen in Lens data as well. KM-ED combination is able to produce most compact clusters, while FCM-MH produces most distinct clusters.

Figures 1(a) and (b) show the FCM and KM-based classification plots of MH and ED distance measures on Iris datasets. Similarly, classification plots have been obtained with Wine and Lens datasets (shown in fig. 2 and 3, respectively).

Table 3. Performance analysis of FCM and KM clustering techniques on **Lens** data

		FCM		**KM**	
Parameters	**Cluster-info**	**MH**	**ED**	**MH**	**ED**
Sensitivity	CL1	0.5	0.5	0.75	0.75
	CL2	1	1	1	1
	CL3	1	1	1	1
	Average	0.8333	0.8333	**0.9166**	**0.9166**
Specificity	CL1	1	1	0.9748	0.9748
	CL2	0.75	0.75	0.875	0.875
	CL3	0.75	0.75	0.875	0.875
	Average	0.8333	0.8333	**0.9082**	**0.9082**
Precision	CL1	0.33	0.33	0.6	0.6
	CL2	1	1	1	1
	CL3	1	1	1	1
	Average	0.7766	0.7766	**0.8666**	**0.8666**
Accuracy	*Average*	0.833	0.833	**0.917**	**0.917**
Time	*Average*	0.0429	0.0429	**0.0296**	**0.0296**
Inter cluster Distance	CL1	1.4889	1.3609	2.2333	**1.2871**
	CL2	1.6174	1.0599	1.7375	**0.9521**
	CL3	2.5152	1.3514	1.8958	**1.1197**
Intra cluster Distance	CL1-CL2	**4.7358**	3.0057	4.9583	3.0016
	CL2-CL3	**4.3081**	2.5404	5.2583	2.8585
	CL1-CL3	**5.1177**	2.8430	3.8958	2.3852

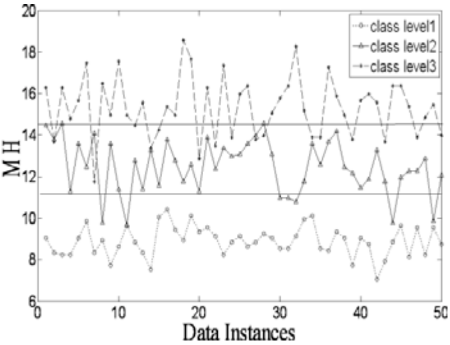

Fig. 1a. Classification of **Iris** data using **FCM** algorithm and **MH** distance

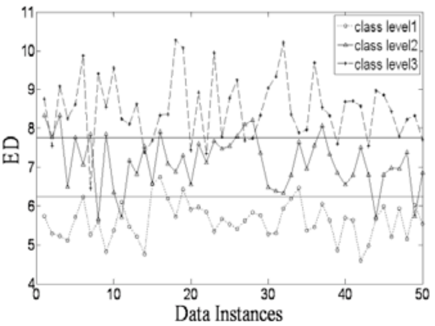

Fig. 1b. Classification of **Iris** data using **FCM** algorithm and **ED** distance

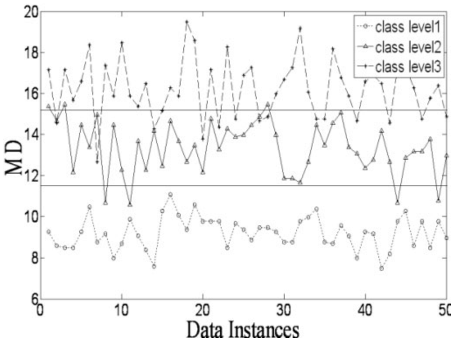

Fig. 2a. Classification of **Iris** data using **KM** algorithm and **MH** distance

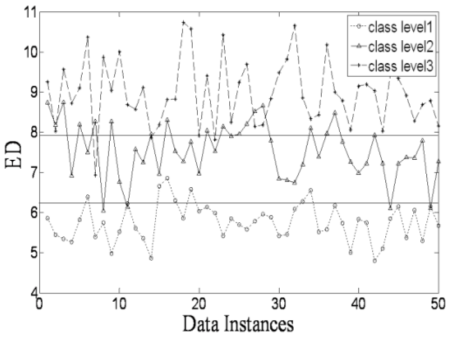

Fig. 2b. Classification of **Iris** data using **KM** algorithm and **ED** distance

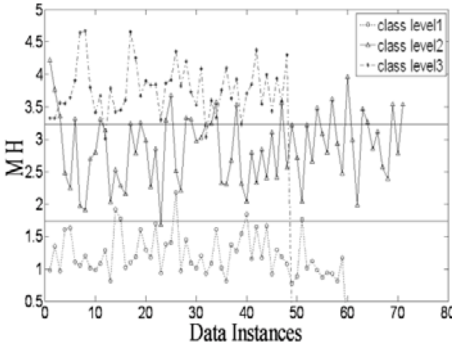

Fig. 3a. Classification of **Wine** data using **FCM** algorithm and **MH** distance

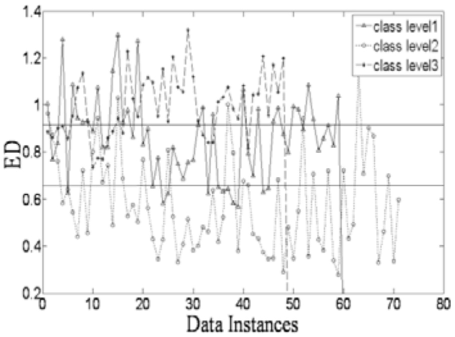

Fig. 3b. Classification of **Wine** data using **FCM** algorithm and **ED** distance

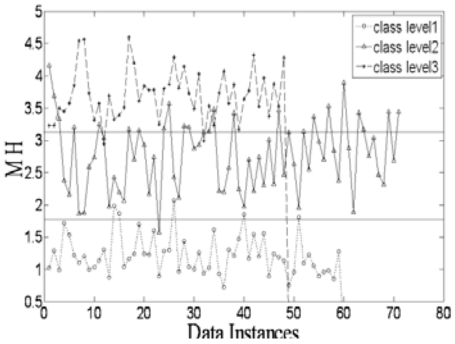

Fig. 4a. Classification of **Wine** data using **KM** algorithm and **MH** distance

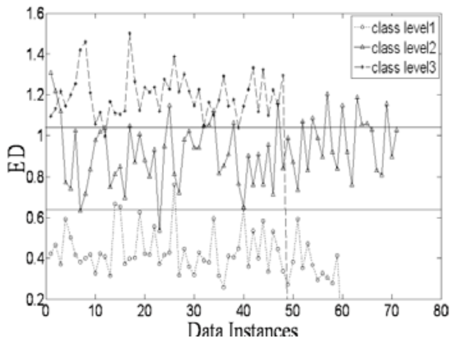

Fig. 4b. Classification of **Wine** data using **KM** algorithm and **ED** distance

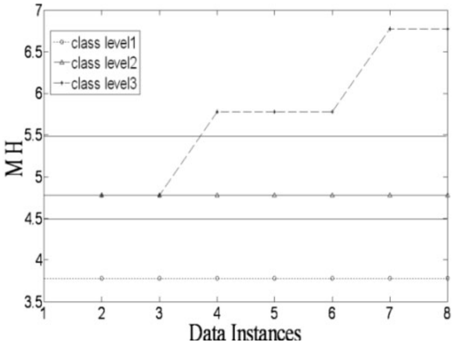

Fig. 5a. Classification of **Lens** data using **FCM** algorithm and **MH** distance

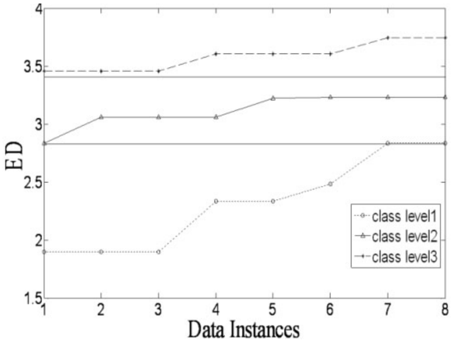

Fig. 5b. Classification of **Lens** data using **FCM** algorithm and **ED** distance

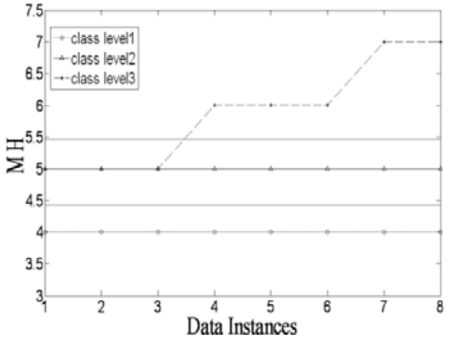

Fig. 6a. Classification of **Lens** data using **KM** algorithm and **MH** distance

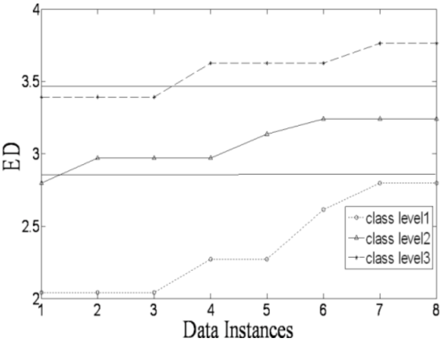

Fig. 6b. Classification of **Lens** data using **KM** algorithm and **ED** distance

Computational time has finally been computed. Figure 7(a), (b), and (c) shows the respective plots for MH distance as it is obvious that the amount of computation in MH is much less that the ED.

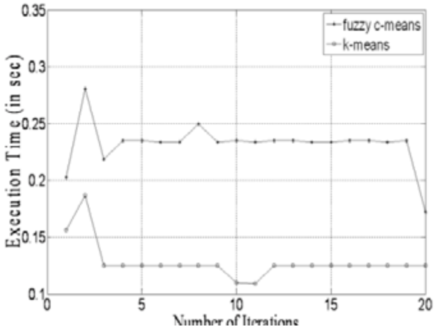

Fig. 7a. Run time plots on **Iris** data

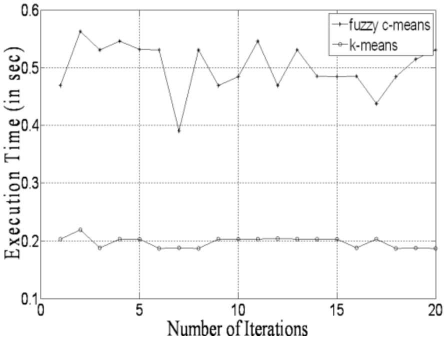

Fig. 7b. Run time plots on **Wine** data

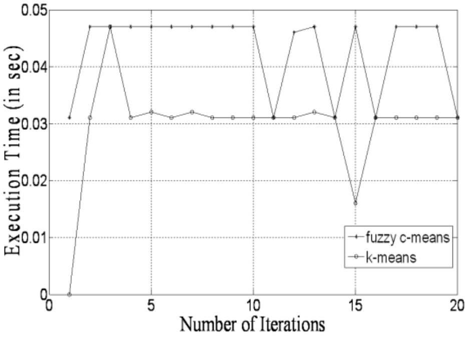

Fig. 7c. Run time plots on **Lens** data

4 Conclusions and Future Work

The results reveal that on Iris data higher precision values for clustering have been obtained with both MH and ED. With FCM-MH combination, more distinct clusters are produced. On the other hand, higher accuracy could be revealed with KM-MH and KM-ED combinations. With the same combination, KM is able to produce high quality clusters with minimum run time. Finally, KM-ED combination is able to yield most compact clusters. In the Wine data, with FCM-MH combination produces most compact and distinct clusters with greater accuracy and minimum time. In the Lens data, FCM-MH can produce the most distinct clusters. For the other parameters, KM performs better than the FCM algorithm. From the results, it is observed that, overall, KM outperforms FCM.

In future, the algorithms could be implemented on real-life clinical data, which are much subjective in nature. Therefore, it would be challenging to choose the right clustering-distance measure approach. Currently the authors are working on this topic.

References

[1] Bezdek, J.C.: Fuzzy mathematics in pattern classification. Applied Mathematics Centre, Cornell University, Ithaca. PhD thesis (1973)

[2] Keller, J., Gary, M.R., Givens, J.A.: A fuzzy k-nearest neighbor algorithm. IEEE Tr. Syst. Man Cyber. 15(4), 580–585 (1985)

[3] Yao, J., Dash, M., Tan, S.T., Liu, H.: Entropy-based fuzzy clustering and fuzzy modeling. Fuzzy Sets and Systems 113, 381–388 (2000)

[4] MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, CA, pp. 281–297 (1967)

[5] Sibson, R.: SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal (British Computer Society) 16(1), 364–366 (1973)

[6] Chattopadhyay, S., Pratihar, D.K., De Sarkar, S.C.: A comparative study of fuzzy C-means algorithm and entropy-based fuzzy clustering algorithm. Computing and Informatics 30(4), 701–720 (2011)

[7] `http://archive.ics.uci.edu/ml/` (Online; last accessed on December 23, 2011)

[8] Han, J., Kamber, M. (eds.): Data Mining Concepts and Techniques, 2nd edn. Elsevier, San Fransisco (2006)