Improving the selection of features through Fuzzy C-means Clustering Technique

1. Poornima G Patil
Department of MCA
VTU Centre for Post Graduate Studies
Bangalore, India
Poornima_g_patil@yahoo.com

2. Dr. Ravindra S Hegadi
School of Computational Sciences
Solapur University
Solapur, Maharashtra
rshegadi@gmail.com

***Abstract*—Clustering is an unsupervised learning method which can separate the data into distinct clusters. The proposed study aims to improve the selection of features through Fuzzy C - means clustering technique. The study is conducted on the images of handwritten signatures form the standard GPDS database. The images after the preprocessing steps of putting bounding rectangles over the signature area, size normalization and thinning are subjected to wavelet decomposition and the resulting detail and approximation coefficients are subjected to principal component analysis for ten levels. The three principal components of detail coefficients and approximation coefficients are used as features in the feature vector. Fuzzy-C means clustering is employed on the features and iteratively the optimization of objective function is achieved and new cluster centers are formed with respect to different features. The new cluster centers obtained for certain features which have good separating ability indicate the selection of those features for training the classifier and achieving better classification results.**

***Keywords*—wavelet decomposition; principal component; Fuzzy C-means Clustering, cluster centers;**

I. INTRODUCTION

The proposed work uses the principal components of detail and approximation coefficients for handwritten signature images and further analyze them to determine which combination of features will be able to distinguish between the signatures of one person from another. Fuzzy C-means Clustering is used for the study. Data clustering is the process of assigning the data points into classes or clusters so that data points of the same class are as similar as possible, and data points in different classes are as dissimilar as possible. Different measures of similarity are used to place the data points into classes. Similarity measures like distance, connectivity and intensity are generally used for clustering.

Basically there are two type of clustering a) Hard clustering and b) Soft clustering. Hard clustering divides the data into distinct clusters, where each data point is assigned to exactly one cluster .Soft clustering or Fuzzy Clustering data points can be assigned to more than one cluster, and associated with each point is a set of membership levels. The membership levels for a data point indicate the strength of the association between that data point and a particular cluster. Fuzzy clustering assigns these membership levels to each data point and then assign the data points to one or more clusters based on them.

Fuzzy C-means (FCM) Algorithm is one of the most commonly used fuzzy clustering algorithms. This clustering technique aims to partition a finite collection of data points $X = \{x_1, ..., x_n\}$ into c fuzzy clusters with respect to some given criterion. Given a finite set of data points the algorithm generates a set of cluster centers $C = \{c_1, ..., c_c\}$ and a partition matrix $W = w_{i,j} \in [0, 1],\ i = 1, ..., n,\ j = 1, ..., c$

Where each element w_{ij} indicates the degree to which data point x_i belongs to cluster c_j. FCM algorithm aims to minimize an objective function and the standard function is:

$$w_k(x) = \frac{1}{\sum_j \left(\frac{d(\text{center}_k, x)}{d(\text{center}_j, x)}\right)^{2/(m-1)}}. \quad (1)$$

Where m is a fuzzifier which determines the level of cluster fuzziness. The selection of a large m results in smaller memberships w_{ij} and hence, fuzzier clusters and in the limit of m = 1, the memberships w_{ij} converge to 0 or 1, which results in a crisp partitioning. The value of m is commonly set to 2 in the absence of lack of domain knowledge or experimentation.

Every data point has a degree of association to clusters in Fuzzy clustering as in fuzzy logic rather than just belonging completely to one cluster. Hence, points on the edge of a cluster may have lesser degree of association with the cluster than the points in the center of cluster.

Each data point x has a set of coefficients giving the degree of being in the kth cluster $w_k(x)$. With fuzzy C-means, the centroid of a cluster is calculated which is the mean of all points, weighted by their degree of belonging to the cluster and this is done iteratively.

$$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}. \quad (2)$$

The degree of association $w_k(x)$ is related inversely to the distance from x to the cluster center as calculated in the previous iteration. The fuzzifier parameter m controls how much weight is given to the closest center. The fuzzy c-means algorithm has the following steps.

1. Choose the suitable number of clusters.
2. Initially each point is assigned the coefficients for being in the clusters.

3. Repeat till the algorithm converges. The convergence means the change in the coefficients between two iterations is not more than the given tolerance or sensitivity threshold.

- Equation (2) is used to compute the centroid for each cluster.
- For each data point, the coefficients of being in the clusters are computed using the Equation (1).

The intra-cluster variance is minimized by the algorithm. Fuzzy C-means has been widely used for clustering of objects present in images.

II. WAVELET RELATED CONCEPTS

A wavelet is a waveform having an average value of zero and lasting for a limited duration of time. They differ from sinusoids in the fact that a wavelet has a beginning and an end and sinusoids extend form minus infinity to plus infinity. Wavelets are capable of representing and analysing the signals at more than one resolution which is called as multi-resolution ability. The multi-resolution analysis has an advantage that the features which go undetected at one resolution may be detected at other resolutions. Wavelets are capable of analysing both stationary and non-stationary signals.

The wavelet is stretched and shifted to correlate with any event which is of interest so that the frequency and time of the event can be exactly measured. As a result of a signal being decomposed by a wavelet two types of coefficients namely detail coefficients and approximation coefficients are obtained. Longer portion of the signal is compared with the wavelet when the wavelet is stretched and they represent the low frequency components which are nothing but slowly varying parts of the signal. Smaller portion of the signal is compared to the wavelet when a wavelet is shrunk and they represent high frequency components which are the rapidly changing parts of the signal. Two types of transforms Continuous and Discrete Wavelet Transforms are possible.

A. Discrete Wavelet Transform

The signal is analyzed at dyadic scales and positions which are powers of two resulting in an accurate analysis instead of analyzing the signal at each scale and position which is expensive.

B. Continuous Wavelet Transform

This transform is continuous meaning the signal is analysed fully by the wavelet. The scaled and shifted wavelet is multiplied with the signal and summed for the entire duration of the signal.

III. RELATED WORK

Many research works have been conducted using Fuzzy C-means clustering technique. A study where the comparison between K-means and Fuzzy C-means algorithms is conducted reports that both K-means and Fuzzy C-means have performed well but, K-means algorithm outperforms Fuzzy C-means in terms of speed. The Fuzzy C-means algorithm using Manhattan distance produces most compact clusters and K-means with Euclidean distance yields most distinct clusters [1]. A paper compares K-means and Fuzzy C-means during the segmentation of color images and it is reported that K-means clustering produces higher accuracy and requires less computation. Although Fuzzy C-means clustering produces results closer to that of k-means requires more computation time because of the fuzzy measure calculations involved in the algorithm [2]. Another study where in the segmentation of images is done using K-means and Fuzzy C-means states that simulation results of both the algorithms are same but the execution time for K-means is 0.302351 seconds and for fuzzy C-means it is 4.40076 seconds [3].

The comparison of Fuzzy C-means and an entropy based Fuzzy C-means reports that the Fuzzy C-means clustering algorithm's performance is better than entropy based Fuzzy C-means algorithm in case of a certain dataset whereas the entropy based Fuzzy C-means algorithm's performance is better than Fuzzy C-means on other two data sets. Entropy based Fuzzy C-means is found to yield more distinct and compact clusters [4]. An enhanced Fuzzy C-means clustering is used for segmenting highly corrupted images and the algorithm uses the spatial and gray level information of the neighborhood of a pixel in a fuzzy way to cluster the corrupted pixel correctly. This algorithm is said to guarantee noise insensitiveness and image detail preservation. It is able to cluster the noisy images corrupted up to 60% [5].

A survey of Fuzzy C-means algorithm reveals that the results of the Fuzzy C– means clustering are more accurate in comparison with the results of the Hard C – means clustering since Fuzzy algorithm allows gradual memberships of data points to the clusters [6]. A fuzzy segmentation algorithm uses a suppressed fuzzy -means clustering (FSSC) algorithm, which directly considers object similar surface variations and perceptually selects the threshold within the range of human visual perception. The qualitative and quantitative results of FSSC algorithm are better than Fuzzy C-means clustering and probabilistic C-means (PCM) for many different images. The results are said to be highly dependent on the features used and the type of the objects in a particular image [7].

An application of the Hierarchical clustering and Fuzzy C-Means clustering method on the analysis of non-preprocessed Fourier-transform infrared spectroscopy (FTIR) data for cancer diagnosis reports that the Fuzzy C-means algorithm performs better than the Hierarchical clustering [8]. The paper compares the results of a numerical comparison of two versions of the Fuzzy C-means (FCM) clustering algorithms. An approximate Fuzzy C-means (AFCM) implementation is proposed where the exact variants in the FCM equation are replaced by the integer-valued or real-valued estimates. This approximation helps AFCM to use a lookup table approach for computing Euclidean distances and for exponentiation. The result of this implementation is that the CPU takes one sixth of the time required for a literal implementation on a nine-band digital image [9]. An Anomaly detection system for network flow employs a method combined with the average information entropy, support vector machine and fuzzy genetic algorithm. These hybrid algorithms are more accurate in classification [10].

A comparison between two clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms based on the efficiency of the clustering output. The factors like the numbers of data points as well as the number of clusters are the factors upon which the behaviour patterns of both the algorithms are

analyzed. FCM is said to produces results closer to K-Means clustering and requires more computation time than K-Means clustering [11]. In order to overcome the noise sensitiveness of conventional fuzzy c-means (FCM) clustering algorithm, a novel extended FCM algorithm for image segmentation has been proposed where the objective function of the standard FCM algorithm is modified with a penalty term that considers the influence of the neighboring pixels on the center pixels. The penalty term behaves as a regularizer in this algorithm which is inspired by the neighborhood expectation maximization algorithm. The experimental results on segmentation of synthetic and real images report that the proposed algorithm is effective and robust [12].

IV. PROPOSED SYSTEM

The system has been developed in MATLAB software. The handwritten signature images are from the standard GPDS database. For each signer, 23 genuine signature images are decomposed using wavelet decomposition for ten levels generating detail and approximation wavelet coefficients. These coefficients are subjected to principal component

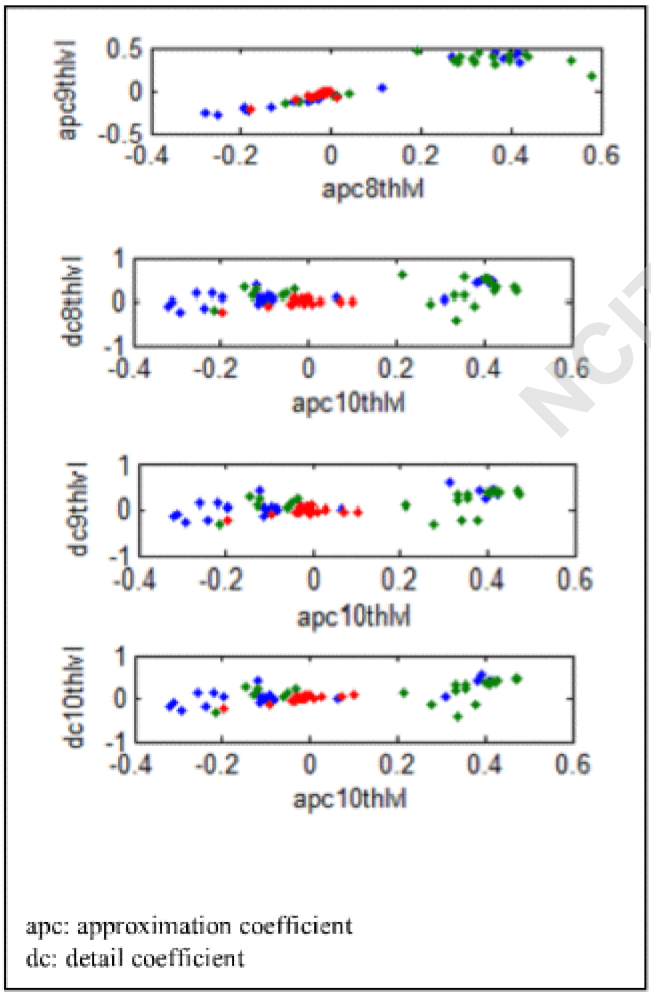

Fig. 1. Sample plots of data for different combination of feature pairs

analysis. Among the first ten principal components of both approximation and detail coefficients of each signature only the eighth, ninth and tenth principal components of approximation and detail are used in the feature vector. For each signer six different features extracted from the 23 genuine signatures make the feature vector. Each signer has 138 (23*6) feature values in the feature vector. The feature vector of each signer are separated from others. Since there are six different features, 15 combinations of 6 feature values are formed. The data is plotted for each pair with one feature on the x-axis and another on the y axis. The sample plots are plotted as shown in the Fig. 1. In the figure 'apc' stands for approximation coefficient and 'dc' stands for detail coefficient. Label on X and Y axes have apc or dc along with the level name. For example apc10thlvl means approximation coefficient at 10^{th} level. The sample plots do not have any clarity with respect to any combination of features pair. In order to gain clarity with respect to combinations of feature values, Fuzzy C-means clustering is used.

The Fuzzy C-means clustering uses the following four parameters. They are the number of clusters, an exponent value, maximum number of iterations and the sensitivity threshold. The exponent value of 2.0 is chosen and sensitivity threshold or minimum improvement parameter is set the value of 1e-6. Maximum number of iterations is chosen as 100.

After the feature extraction stage, Fuzzy C-means clustering (FCM) is used where in the dataset is divided into clusters with every data point in the dataset belonging to every cluster with a certain degree. A certain data point that lies closer to the center of a cluster will have a high degree of association or membership to that cluster and another data point that lies far away from the center of a cluster will have a low degree of belonging or membership to that cluster.

Fuzzy C-means clustering starts with an initial guess for the cluster centers the purpose of which is to mark the mean location of each cluster. Generally the initial guess values of cluster centers are incorrect. FCM assigns to every data point a membership grade with respect to each cluster. The algorithm updates the cluster centers and the membership grades for each data point iteratively. Eventually the cluster centers are moved to the right location within a data set. The iterative updation is based on minimizing an objective function which represents the distance from any given data point to a cluster center weighted by that data point's membership grade.

The algorithm initially starts with some cluster centers. For each iteration, for each combination of features pair new cluster centers are updated and also the membership grade of each data point is updated. This process continues till the objective function converges meaning the difference of value of the objective function in the current iteration and the previous iteration is less than the sensitivity threshold (minimum improvement) or the maximum number of iterations are over. The plots of data for different combinations of feature pairs with the new cluster centers represented in bold letters are shown in the figure 2. The plot of combination of features for example approximation coefficients at 10^{th} level (apc10thlvl) against detail coefficients at 8^{th},9^{th} and 10^{th} level (dc8thlvl, dc9thlvl, dc10thlvl) (the last three plots in figure 2) show distinct clusters indicating the fact that these features when given higher weights can yield higher accuracy during classification process. The new cluster centers are shown in bold on the respective figures. The x and y co-ordinate values of the new clusters centers 1, 2 and 3 are shown in the table I for all 15 combinations of feature pairs.

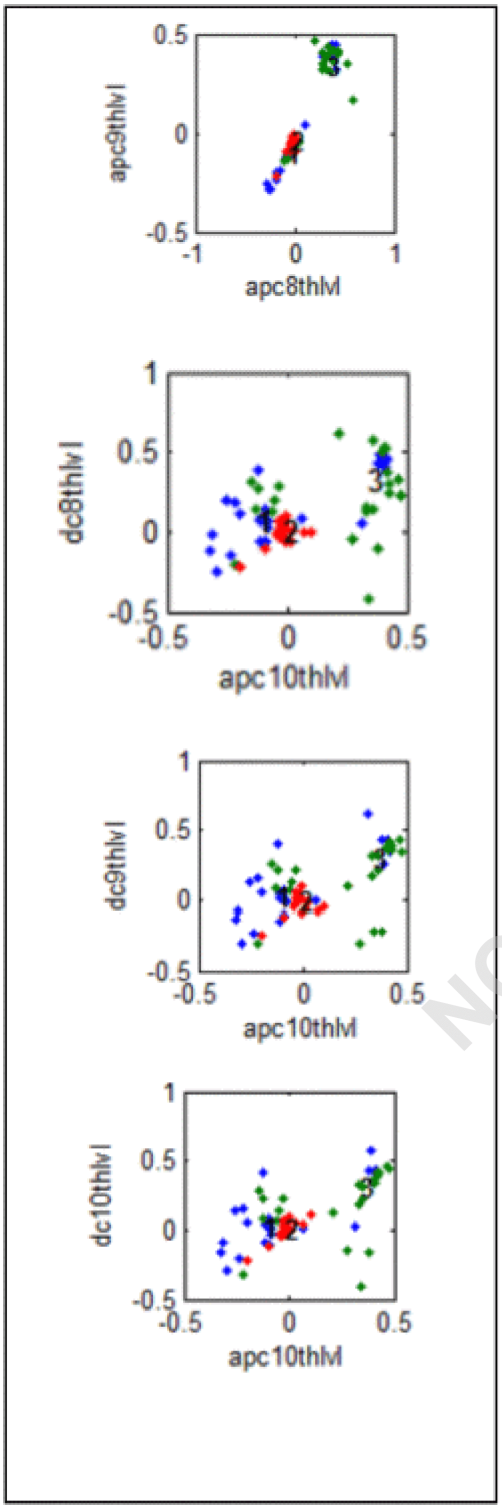

Fig. 2. New cluster centers for certain combinations of feature pairs

V. RESULTS AND CONCLUSION

New cluster centers formed after the application of Fuzzy C-means clustering is shown in the Table I. For the combinations 10, 11 and 12 (approximation coefficient at 10^{th} level versus detail coefficient at 8^{th} , 9^{th} and 10^{th} level have the three cluster centers which are quite apart when compared to other combinations.

Combination	Cluster	X co-ordinate	Y Co-ordinate
1	1	-0.0223	-0.0416
	2	-0.0705	-0.1102
	3	0.3323	0.3281
2	1	-0.0223	-0.0164
	2	-0.0705	- 0 .1161
	3	0.3323	0.3383
3	1	-0.0223	-0.0014
	2	-0.0705	0.0615
	3	0.3323	0.3146
4	1	-0.0223	-0.0182
	2	-0.0705	0.0124
	3	0.3323	0.2876
5	1	-0.0223	0.0040
	2	-0.0705	- 0.0166
	3	0.3323	0.3047
6	1	-0.041	-0.0164
	2	-0.1102	-0.1161
	3	0.3281	.3383
7	1	-0.0416	-0.0014
	2	-0.1102	-.0615
	3	0.3281	0.3146
8	1	-0.0416	-0.0182
	2	-0.1102	0.0124
	3	0.3281	0.2876
9	1	-0.0416	0.0040
	2	-0.1102	0.0166
	3	0.3281	0.3047
10	1	-0.0164	-0 .0014
	2	-0.1161	0.0615
	3	0.3383	0.3146
11	1	-0.016	-0.0182
	2	-0.1161	0.0124
	3	0.3383	0.2876
12	1	-0.0164	0.0040
	2	-0.1161	0.0166
	3	0.3383	0.3047
13	1	-0.0014	-0.0182
	2	0.0615	0.0124
	3	0.3146	0.2876
14	1	-0.0014	0.0040
	2	0.0615	0.0166
	3	0.3146	0.3047
15	1	-0.0182	0.0040
	2	0.0124	0.0166
	3	0.2876	0.3047

Table I. New Cluster centers formed after Fuzzy C-means Clustering

Therefore the features in these combinations have higher discriminating power and hence can be selected for further analysis to probe that they can produce more accurate results if used for classification of signature samples. Fuzzy C-means clustering can help select discriminating features which can

reduce the dimension of feature space and also yield better classification results.

REFERENCES

[1] Sandeep Panda, Sanat Sahu, Pradeep Jena, Subhagata Chattopadhyay, " Comparing Fuzzy-C Means and K-Means Clustering Techniques: A Comprehensive Study," Proceedings of the Second International Conference on Computer Science, Engineering and Applications (ICCSEA 2012), Volume 1, pp 451-460, May 25-27, 2012, New Delhi, India.

[2] Bharati R.Jipkate, V.V.Gohokar, "A Comparative Analysis of Fuzzy C-Means Clustering and K Means Clustering Algorithms ," IJCER, Vol. 2, Issue No.3 , pp 737-739 , May-June 2012.

[3] Mamta Bharadwaj, Ankita Walia, Hemant Tulsani, " Comparative Research on Fuzzy C-Means and K-Means Clustering Segmentation , " Advanced International Journal of Computer Applications and Information Technology, ISSN: 2278-7720, Vol. 3, Issue II, page 44,Aug-September 2013.

[4] Subhagata Chattopadhyay, Dilip Kumar Pratihar," A Comparitive Study of Fuzzy C-means Algorithm and Entropy Based Fuzzy Clustering Algorithms," Computing and Informatics, Vol. 30, 701–720 2011.

[5] N.MymoonZuviria , M..Deepa, "Robust Fuzzy Neighborhood Based C Means Algorithm for Image Clustering," International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 , volume 3, Issue 3, March 2013.

[6] Makhalova Elena, " Fuzzy C-means Clustering in Matlab," The 7th International Days of Statistics and Economics, Prague, September 19-21, 2013.

[7] M. Ameer Ali, Gour C Karmakar and Laurence S Dooley, " Fuzzy Image Segmentation using Suppressed Fuzzy C-Means Clustering," International Conference on Computer and Information Technology, Dhaka, Bangladesh ,26th - 28th December 2004.

[8] Xiao Ying Wang, Jon Garibaldi, Turhan Ozen, "Application of the Fuzzy C-Means Clustering Method on the Analysis of non Preprocessed FTIR Data for Cancer Diagnosis," Department of Computer Science and Information Technology, The University of Nottingham, United Kingdom.

[9] Robert L Cannon, Jitendra V Dave, James C Bezdek, "Efficient Implementation of the Fuzzy c-Means Clusteng Algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No.2, March 1986.

[10] Linquan Xie, Ying Wang, Liping Chen, Guangxue Yue, "An Anomaly Detection Method Based on Fuzzy C-means Clustering Algorithm," Proceedings of the Second International Symposium on Networking and Network Security, ISBN 978-952-5726-09-1, 2010,

[11] Soumi Ghosh, Sanjay Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[12] Yong Yang, Shuying Huang," Image Segmentation By Fuzzy Computing and Informatics," Vol. 26, 17-31,2007.

[13] J.F.Vargas, M.A.Ferrer, C.M.Travieso, and J.B.Alonso, "Off- line signature verification based on grey level information using text features", 44(2):375–385, 2011.