Improved Fuzzy C-Means clustering with application for patients of Iron Deficiency Anemia discrimination

By

SAHAR ADEL RAAFAT

Assistant Professor, faculty of Commerce,

Zagazig University, Benha Branch

ABSTRACT :

The Fuzzy C-Means Algorithms (FCMA) was applied for the discrimination between groups with Iron deficiency Anemia by artificial vision platelet counts of patient with certain measure. The aim of this study is to cluster, by FCMA, the measured learning and test data into 3 groups. The FCMA process was improved by the introduction of non-random initialisation of the cluster centers. The improved FCMA is a faster algorithms than the standard FCMA, it enables the investigator to test different feature vector representations quickly, which otherwise would have been impractical. In addition the Mahalanobis distance was used, instead of Euclidean for measuring the proximity of a pattern to a cluster distance. Furthermore, a classification approach with a reject threshold was investigated for increasing the classification performances. This was achieved by assigning the patients which were lying in the fuzzy boundaries between the available classes to reject class. The initialisation method was introduced in terms of the computation time and the percentage of correct recognition, a comparison study between Mahalanobis

distance and Euclidean was carried for measure the error of classification. It was shown that the use of the Mahalanobis distance improved the performance in comparison to the Euclidean distance, the Mahalanobis distance allowed 95.7%, 95% of the learning and test sets to be correctly identified and also allowed an improvement in the correct recognition rates. A further experiment showed that the value of the stopping criterion ε had little influence on the recognition rates, but it had a large influence on the computation time of FCMA.

1. Introduction :

Classification using different types of clustering techniques is an efficient tool for the statistics. Fuzzy clustering techniques allow an object to have membership in more than one cluster. The need for fuzzy clustering occurs when objects tend to occupy positions in the feature space between more clearly defined clusters.

The classification techniques can be grouped into 2 main categories : supervised and non-supervised classification. In supervised classification, the qualitative group of each learning pattern has to be known during the learning process. On the contrary, the non-supervised learning approach does not require knowledge of the labels of learning patterns **Chtioui, *et al.* (1996).** However, their computation time is relatively high due to a large number of iterations needed before the algorithm converges.

Non-supervised learning approach is attractive because the a priori knowledge of the labels of the learning patterns is not always

available, so its purpose is to partition a set of patterns into consistent groups (clusters). Patterns of a given cluster have to share common properties which distinguish them from the other clusters.

A supervised pattern recognition technique is relatively fast because it does not need to iterate; however, it cannot be applied to a problem unless training data or relevant knowledge are available.

The two types techniques are complementary. For example, non-supervised clustering can be used to produce classification information needed by a supervised pattern recognition technique . For achieving non-supervised classification, there are two ways such as hierarchical and non-hierarchical methods, the first methods, such as the reciprocal neighbours method **John & Reza (1991),** group data in ascending or descending manner, by constructing a tree, but they are not suitable for clustering large data sets **Spath (1980)**. The second methods are based on the optimisation of an objective function, these methods are known as clustering method such as the C-Means, the ISO data **Dunn (1974)** and the Kohonen self-organising feature map **Kohonen (1989),** these methods don't take into account the fact that boundaries between classes may fuzzy. Fuzzy sets were introduced in order to represent vagueness **Bezdek (1981).**

It is appealing to apply fuzzy sets in clustering problems, because they are able to consistently describe some of uncertainties that often exist in data.

The Fuzzy C-Means Clustering Algorithm (FCMA) was proposed by **Bezdek *et al.*, (1987),** and **Chtioui *et al.*, (1997)** it is a fuzzy logic-based extension to the hard C-means clustering algorithm. Both of them are unsupervised techniques. The FCMA proposed by **Chtioui *et al.*, (1997)** was improved by adding same proposed steps using a reformulation of the FCM algorithm in terms of matrix algebra which makes it efficient for such data sets.

The foundation of the FCM algorithm is to optimize an objective function that measures the compactness and the separation of clusters formed.

FCMA is based on fuzzy sets which makes it possible to assign a membership function to each pattern, this function evaluates the degree of membership of a pattern to each cluster. The standard FCMA has usually been applied with random locations of the initial cluster centers, so, the performances of FCMA greatly depend on the initial location of the cluster centers. For solving this problem the algorithm has to be applied several times with different initialisations of the cluster centers to select the best performance. This approach is time consuming and a method for finding the optimal initial cluster centers is clearly desirable. Also, the choice of a distance measure between a pattern and a cluster center may play an important role in the correct separation between groups. **Chtioui *et al.*, (1997).**

The need for fuzzy clustering occurs when the number of features is significantly large and FCMA is unnecessarily computer

demanding for such case, and the whole analysis becomes unnecessarily slow and tedious. These considerations have motivated to improve FCMA process by an algorithm which is much faster than the standard FCMA for testing different feature vector representations quickly. In most studies, the Euclidean distance has been used. The Euclidean distance does not take into account the underlying structures of clusters, but the Mahalanbis distance was used as a measure of proximity of a pattern to a cluster.

In order to take into account the real data structure in the feature space, the Mahalanobis distance was used instead of the Euclidean distance. FCMA was also applied with a reject threshold patterns which have approximately the same degree of membership to all the available classes were rejected i.e. not classified. A range of experiments were performed in order to investigate the effect of the rejection rates on the classification performance.

FCMA was applied on a medical problem which was to asses different rates of IDA (Iron Deficiency Anemia) using artificial vision.

2. Theory :

2.1. The theory and general principles of FCM clustering algorithm:

Chtioui *et al.*, (1996) defined the algorithm as follows :

Let n patterns be clustered into C groups, where each pattern is a P-dimensional vector, and noted x_i, $i \in [1, 2, \ldots, n]$.

Let W : be a membership matrix (n rows and c columns).

All the elements of W are numbers between zero and one.

The element w_{ij} of W : represents the membership degree of the pattern i to the class j. The sum of the membership degrees of a pattern to all the available classes is one. The cluster centers are grouped in a matrix Z (c row and p columns). Each raw of this matrix, noted z_i, refers to the cluster center of class i.

Bezdek *et al.*, (1987) introduced the steps where FCMA adjusts the elements of Z matrix as follows :

(1) Set the iteration number K to 0, and choose an initial cluster centers matrix Z (k).

(2) Assess the matrix of membership degrees according to :

$$w_{ij}(k) \frac{[1/(d^2(x_i, z_j)]^{1/(m-1)}}{\Sigma_{h=1}^{c} [1/(d^2(x_i, z_h))]^{1/(m-1)}}$$

$$\forall\, i \in (1, 2, \ldots, n) \text{ and } \forall\, j \in (1, 2, \ldots, c) \qquad (1)$$

The parameter m is a scalar larger than 1 which controls the fuzziness of FCMA.

(3) Update the matrix of cluster centers :

$$z_j(k+1) = \frac{\Sigma_{i=1}^{n} w_{ij}(k)^m x_i}{\Sigma_{i=1}^{n} w_{ij}(k)^m} \qquad (2)$$

This is simply a weighted average of the data.

(4) If $\| Z(k) - Z(k-1) \| < \varepsilon$ stop, otherwise increment k, and go to step 2. where ε is a small positive scalar and called the "stopping criterion". The matrix norm is defined by :

$$\left(\sum_{j=1}^{C} \sum_{l=1}^{p} | z_{jl}(k) - z_{jl}(k-1) |^{\prime} \right)^{1/v} \qquad (3)$$

2.2. What is the distance measure :

The assessment of the memberships matrix requires the definition of a distance measure between two patterns. The Euclidean distance is normally used. This metric makes the assumption that the shape of each cluster is hyperspheroidal, which is not usually true.

A much used distance measure is the Minkowski distance between object i and j, **Alsberg (1994)** :

$$d_{ij} = \left(\sum_{k=1}^{M} | x_{ik} - x_{jk} |^{w} \right)^{1/w} \qquad i, j \in [1, \ldots, N) \qquad (4)$$

Where w $\in$ [1, 2, 3, ..] and x_{ik}, x_{jk} are elements in the Matrix X.

Alsberg (1994) presented an extension of the formula for the Euclidean distance matrix presented by **Alsberg and Esbensen (1992)** as follows :

Let D be the matrix containing the distances between the row vectors in X. The squared distances can be written as :

$$d_{ij}^{2} = \left(\sum_{k=1}^{M} (x_{ik} - x_{jk})^{2} \right. \qquad (5)$$

$$\text{Then, } D^{2} = X^{2} J^{T} - 2 X X^{T} + J [X^{2}]^{T} \qquad (6)$$

Where J denotes the 1_{ik} and 1_{jk} matrix elements of the same matrix containing ones.

This equation is more efficiency formulated as :

$$D^{2} = M + M^{T} - 2R \qquad (7)$$

$R = XX^{T}$, $M = 1_{[N\times 1]}$ diag $(R)^{T}$ is not symmetric.

The Mahalanobis distance is more appropriate for describing the real data structure. It was used to measure the distance between a pattern x_i and a cluster center z_i according to :

$$d^2 (x_i, z_j) = (x_i - z_j) F_j^{-1} (x_i - z_j)^t \quad (8)$$

$(x_i - z_j)^t$ = transpose of $(x_i - z_j)$

F_j = Fuzzy variance – covariance matrix for the cluster j, and is defined by :

$$F_j = \frac{\sum_{i=1}^{n} w_{ij} (x_i - z_j)^t (x_i - z_j)}{\sum_{i=1}^{n} w_{ij}} \quad (9)$$

In the FCM algorithm, the distances between objects and cluster centers are calculated, not between the objects themselves. This gives rise to an asymmetric distance matrix.

$$q_{ic}^2 = \sum_{k=1}^{M} (x_{ik} - v_{zk})^2 \quad (10)$$

$$= \sum_{k=1}^{M} (x_{ik}^2 l_{zk} + l_{ik} v_{zk}^2 - 2 x_{ik} v_{zk}) \quad (11)$$

q_{ic} : distance between object i and cluster z v_{zk} : an element in the matrix V.

Equation (11) can be written as :

$$Q^2 = A + B^T - 2 L \quad (12)$$

Where : $A = l_{(N\times 1)} \text{diag} (F)^T$

$F = VV^T$, $B = l_{(k\times 1)} \text{diag} (R)^T$, $L = XV^T$

R, L and F are referred to as kernel matrices and have the following dimensions :

$$\text{Dim}(R) = [N \times N],\ \text{Dim}(L) = [N \times K],\ \text{Dim}(F) = [K \times K]$$

The dimensions of A and B^1 are both [N x K]. The main idea of FCM algorithm is to avoid the numerous distance computations for each iteration step and instead update just the smaller kernel matrices in each iteration.

2.3. The improved FCMA :

This paper proposed an improved FCMA which is much faster than the traditional algorithm for data sets in which the number of features is significantly large. The algorithm is constructed by using the proposed steps of FCMA by **Chtioui *et al.*, (1997)** and adding an estimated of the cluster centers as the following steps :

Chtioui *et al.*, (1997) proposed a deterministic initialisation method, which proceeds as follows :

1) Assess the average μ, and the standard deviation σ of the whole data set.
2) Choose the first cluster center as the average point of the whole data set, μ.
3) Choose an additional cluster center at a far location from the average of the data, with a given number of standard deviations, $\mu + \phi\,\sigma$, where ϕ is a scalar defined by the user.
4) Apply FCMA to partition the data. Repeat steps (3) and (4) until the number of cluster centers is equal to the a priori number of defined classes. The main idea behind this initialisation method

is that the $(k+1)^{th}$ cluster center is placed at a location far from the existing k clusters.

* The proposed adding steps :

5) Let U is the fuzzy membership matrix, add the equations as proposed by **Alsberg (1994).**

$$H = p^{I} \text{ diag } (1./(1^{T}p^{I}), \quad L = RH, \quad F = H^{I}L$$

These equations constitute the updating steps of L and K Kernel matrices, this help to complete the matrix R once and not updated. The distance matrix is transformed such that the sum of the membership values for one object satisfy :

$$\sum_{k=1}^{M} u_{ik} = 1 \qquad \forall i$$

So, only entries from the calculated asymmetric distance matrix are needed. The iteration is stopped when the change in :

$$\sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik}^{m} \left\| x_i - v_k \right\|^2 \quad \text{is below } \varepsilon$$

2.4. Classification with reject :

The decision of rejecting patterns is taken during the de-fuzzification step. A simple way to decide the rejection of a pattern is to compare the variance of its degrees of membership to a user-defined parameter called the "reject threshold". For a given pattern, if the variance is lower than the reject threshold, the pattern is rejected. The rate of rejection is greatly dependent of the value of the reject threshold.

3. Case of Study :

(IDA) is one of the most common maladies in human especially females of reproductive age **(Ernest, *et al.*, 2003).** IDA is a cause of reactive thrombocytosis where a moderate increase in platelet count is common but sometimes count may exceed 1,000 x 10^9/L **(Hamdi, *et al.*, 2000).**

Reactive thrombocytosis associated with IDA is mediated through cytokines that released due to the primary event, when the original stimulation stops the platelet counts returns to the reference range. In this study the sample is consisted of 120 women with IDA, patients with platelet count < 400 x 10^9/L, other with platelet count > 450 x 10^9/L, and the normal controls of other healthy women were included in the study., Epo (Erythropoietin), Tpo (Thrombopoietin) and IL-6 (Intereukin-6) were measured.

IDA is the most common cause of anemia in every country of the world, it is the most important cause of microcytic hypochromic anemia in which red cell indices are reduced, serum iron falls, TIBC (Total Iron Binding Capacity) rises, transferrin saturation < 16% and serum ferritin is very low **(Hoffbrand *et al.*, 2001).**

Thrombocytosis in association with IDA is well documented and is an example of reactive thrombocytosis (RT), it is usually transient and subsides when the primary stimulus ceases. In spite of the high platelet count, thrombotic and/or heamorrhagic complications are highly exceptional **(Susumu *et al.*, 2002).** Mechanisms causing RT remains unclear, based on in vitro studies, thrombopoietic cytokines play a role in RT of IDA. Replacement of iron stores results in normalization of platelet counts **(Hamdi *et al.*, 2000).** Thrombopoietin (Tpo) and intereukin-6 (IL-6), the primary cytokines for platelet production are elevated and stimulate RT, other cytokines

as Erythropoietin (Epo), IL-3, IL-11 and, GM-CSF (Granual – Mono cyte) – Colony Stimulating Factor) **Dina *et al.*, (2003).**

The patient sample consisted of 120 women aged as (20-30), (30-40) years with a demonstrable cause of iron deficiency anemia Using FCM, this sample is divided to 3 groups, each group has 2 ages, and is consisted of :

a) Patients with IDA and platelet counts normal or below 400 x 10^9/L.

b) Patients with IDA and platelet counts > 450 x 10^9/L.

c) (Control group) : Normal healthy women.

Source of Data : El-Mataria Teaching Hospital, Hematology Department, Kasr El-Aini Hospital.

Feature measurement: Table (I) gives a set of 15 features that were extracted for patient characterization.

Table (I)

Feature	
ELISA	Enzyme Linked Immunosorbent Assay
Epo	Erythropoietin
Hb	Himoglobeen
HCT	Himotocrete
IL-6	Interleukin-6
Lys	Lysine
Mc Hc	Mean Corposcolor Himoglobeen Concut
MCH	Mean Corpascolor Himoglobeen
MCV	Mean Cell Volume
plt	Platelet
RT	Reactive Thrombocytosis
S. Ferritin	Serum Ferritin
S. Iron	Serum Iron
TIBC	Total Iron Binding Capacity
Tpo	Thrombopoietin

Moreover, texture features were used for Blood Cell histograms. Each of these features is a 2-dimensional vector according to WBC & RBC (White Blood Cell, Red Blood Cell) as shown in Table (II).

Table (II)

Texture features, each of these features is a 2-dimensional vector

Texture feature	
Contrast	The moment of the inertia of the histogram
Energy	Gives the variance of the histogram of the blood cell level differences
Entropy	Measure the homogeneity of the histogram of the blood cell level differences
Kurtosis	Describes the shape of the blood cell level histogram
Mean	Given the mean level of the histogram of the blood cell level differences
Run length distribution	Given the value for a cell with the most linear structure
Run percentages	For homogeneous texture
Skewness	Describes how symmetric the level histogram of the blood cell
Variance	Measures the dispersion of the blood cell level differences

4. Results and Conclusions :

The purpose of this investigation was to analyse the abilities and the limitations of FCMA for Iron deficiency Anemia-discrimination problem. When applying FCMA, it was possible to assign each patient into one of the 3 groups.

The effect of the improved FCMA was investigated and was contrasted with the standard FCMA and the quality of the initialisation method was based on the percentage of correctly classified learning and test patients.

The use of FCMA requires the definition by the user of several parameters : the matrix norm, the fuzziness parameters m and the

stopping criterion ε. Although there is no theoretical basis for optimising m, this parameter was set to 2.0. **Bezdek (1981).**

The performances of FCMA were assessed for different values of the stopping criterion ε. The reject threshold was varied from 10^{-3} to 10^{-1}.

4.1. Efficiency of the FCM and Improved FCM Algorithms :

The most common way of measuring the efficiency of a numerical algorithm is to obtain the number of floating point operations (Flops) used by the algorithm. Knowing the algorithm, it is also possible to construct a Flops formula, which estimates the Flops consumption for the algorithm given the dimensions of the different input matrices and the number of iterations. For the FCM the Flops formulas are expressed in terms of the dimensions of the input matrix X, (Dim (x) = [N x M]), the number of clusters to be analyzed (k), and the number of iterations (I), which varies according to the structure of data set. **Alsberg (1995)** suggested the following Flops formulas :

$$F_1 = IK\,(7\,NM + 5\,KNM + 3\,KN + 4N + M) \qquad (13)$$

To obtain a measure of the efficiency of algorithms, the Flops command in MATLAB is employed. For each matrix expression in an algorithm, the corresponding Flops formula is found. The total Flops is obtained by summing over the different contributions. Only the most important parts of an algorithm are investigated.

In the Flops formula for the standard FCM algorithm in equation (16), M is involved in each iteration I.

The improved FCM algorithm has the following Flops formula suggested by **Alsberg (1994) :**

$$F_2 = NM\,(N+1) + 2KNM + KM\,(K+1) + IK\,(14N + 1)\,KN + 2N^2 + 1) \qquad (14)$$

The computation of R requires NM (N+1) Flops, the computation of L requires 2KNM Flops, and the computation of F

requires KM (K + 1) Flops. There are the initial calculations of the kernel matrices.

By assuming that the number of objects is constant, N = 120. In addition, the number of iterations (I = 30) and the number of clusters (k = 3) are also constants. The number of features varies from 100 to 1500, the Flops computations for the algorithm is presented in Table (III). This table clearly shows that FCM requires significantly more MFlops (Mega Flops) than the improved FCMA.

Table (III)

The MFlops usage of FCM and improved FCMA Based on the Flops equations for the two algorithms.

M	100	300	500	700	900	1100	1300	1500
FCM: F_1	18.3	37.5	57.2	76.6	96.8	115.7	135.1	154.6
Improved FCMA : F_2	0.9	1.4	1.9	2.5	3.1	3.7	4.3	4.9
r	20.3	26.8	30.1	30.64	31.2	31.3	31.4	31.6

The Flops ratio $r = F_1/F_2$ between the two algorithm, the improved FCMA is useful only when $r > 1$.

4.2. Patient examination :

The 3 groups of patients were chosen because they correspond to a typical patient discrimination problem.

Texture features were extracted from each of the 2 blood cells (WBC, RBC). A set of 18 texture features (9 x 2) were measured, each patient was therefore characterized by a 33 dimensional vector (15 features + 18 texture features). The whole data were gathered into a matrix of sizes 120 rows (patients) and 33 columns (features). This matrix was randomly partitioned into 2 matrices corresponding to the

learning and test data. The size of the learning data was 90 rows x 33 columns, and the test data was 30 rows x 33 columns.

By applying the principal component analysis which is a classical statistical method to create new uncorrelated features, and provides a way to represent data in a reduced space with little loss of information. They are ranked in decreasing order of the total variance they hold. The first 5 principal components were selected because they represented a large percentage (96.02%) of the total variance.

Fig. (1) shows a plot of the whole learning patterns in the plane of the first and the second principal components which described 65.72% of the total variance, each point of this figure represented the projection of a particular patient. patients of (a) were coded as •, patients of (b) were coded as ·, and patients of (c) were coded as *.

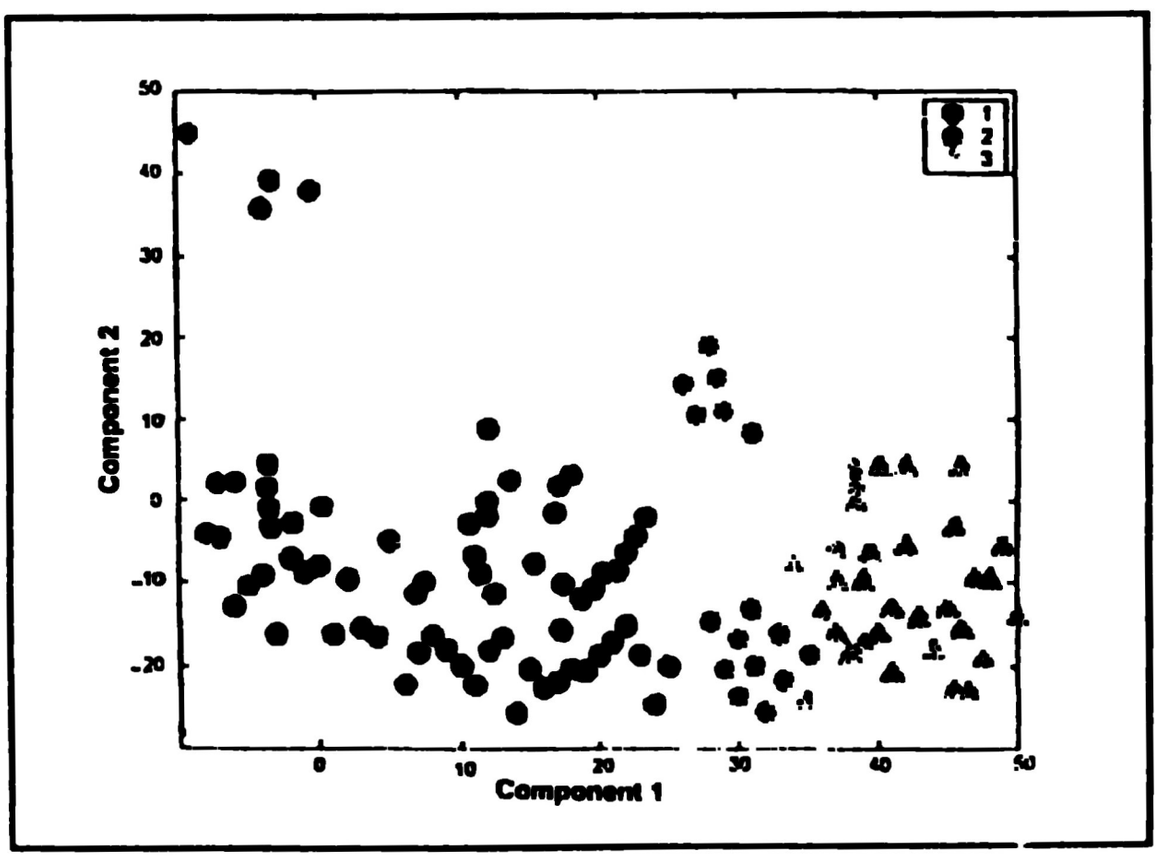

Fig. (1)

The first component is a (roughly) equally weighted sum, or "index", of the variables, this component might be called a general Anemia blood component, the second component represents a contrast value (which is weak) between some features and texture features, it might be called a texture blood component, so that the observations can be almost perfectly classified using only the 1st component, and the 2nd component has no classification power, as shown in Fig (1).

The classification performances were averaged over 10 runs corresponding to 10 random initialisations of the cluster centers. The mean number of iterations of the improved FCM with a random initialisation was 30. The mean correct classification results, with a random initialisation were 95.7% of the training set and 95% of the test. The improved FCMA was better than FCMA with respect to the computation time, and the percentages of correct recognition.

4.3. Effect of the stopping criterion and distance measure :

The parameter ε which had an effect on the classification performances was varied from 10^{-1} to 10^{-5} by steps of one in the base 10 logarithmic scale. For each value of this parameter, FCMA was applied using either Euclidean or the Mahalanobis distance. In each configuration, the classification results of the learning set were assessed. **Fig. (2)** illustrates that the Mahalanobis distance outperformed the Euclidean distance, because the Mahalanobis allowed a decrease of the percentage of misclassified patients by 0.81, when ε was smaller than 10^{-3}. When the stopping criterion was smaller than 10^{-2}, the classification results obtained with the Mahalanbis distance were 95.7% and 95% of correct classifications for training and test sets respectively, as shown in Fig (3).

The first component is a (roughly) equally weighted sum, or "index", of the variables, this component might be called a general Anemia blood component, the second component represents a contrast value (which is weak) between some features and texture features, it might be called a texture blood component, so that the observations can be almost perfectly classified using only the 1st component, as in Fig (1).

The classification performances were averaged over 10 runs corresponding to 10 random initialisations of the cluster centers. The mean number of iterations of the improved FCM with a random initialisation was 30. The mean correct classification results, with a random initialisation were 95.7% of the training set and 95% of the test. The improved FCMA was better than FCMA with respect to the computation time, and the percentages of correct recognition.

4.3. Effect of the stopping criterion and distance measure :

The parameter ε which had an effect on the classification performances was varied from 10^{-1} to 10^{-5} by steps of one in the base 10 logarithmic scale. For each value of this parameter, FCMA was applied using either Euclidean or the Mahalanobis distance. In each configuration, the classification results of the learning set were assessed. Fig. (2) illustrates that the Mahalanobis distance outperformed the Euclidean distance, because the Mahalanobis allowed a decrease of the percentage of misclassified patients by 0.81, when ε was smaller than 10^{-3}. When the stopping criterion was smaller than 10^{-2}, the classification results obtained with the Mahalanbis distance were 95.7% and 95% of correct classifications for training and test sets respectively, as shown in Fig (3).

The first component is a (roughly) equally weighted sum, or "index", of the most of variables, this component might be called a general Anemia blood component, the second component represents a contrast between some features and texture features, it might be called a texture blood component.

The classification performances were averaged over 10 runs corresponding to 10 random initialisations of the cluster centers. The mean number of iterations of the improved FCM with a random initialisation was 30. The mean correct classification results, with a random initialisation were 95.7% of the training set and 95% of the test. The improved FCMA was better than FCMA with respect to the computation time, and the percentages of correct recognition.

4.3. Effect of the stopping criterion and distance measure :

The parameter ε which had an effect on the classification performances was varied from 10^{-1} to 10^{-5} by steps of one in the base 10 logarithmic scale. For each value of this parameter, FCMA was applied using either Euclidean or the Mahalanobis distance. In each configuration, the classification results of the learning set were assessed. **Fig. (2)** illustrates that the Mahalanobis distance outperformed the Euclidean distance, because the Mahalanobis allowed a decrease of the percentage of misclassified patients by 0.81, when ε was smaller than 10^{-3}. When the stopping criterion was smaller than 10^{-2}, the classification results obtained with the Mahalanbis distance were 95.7% and 95% of correct classifications for training and test sets respectively, as shown in **Fig (3).**

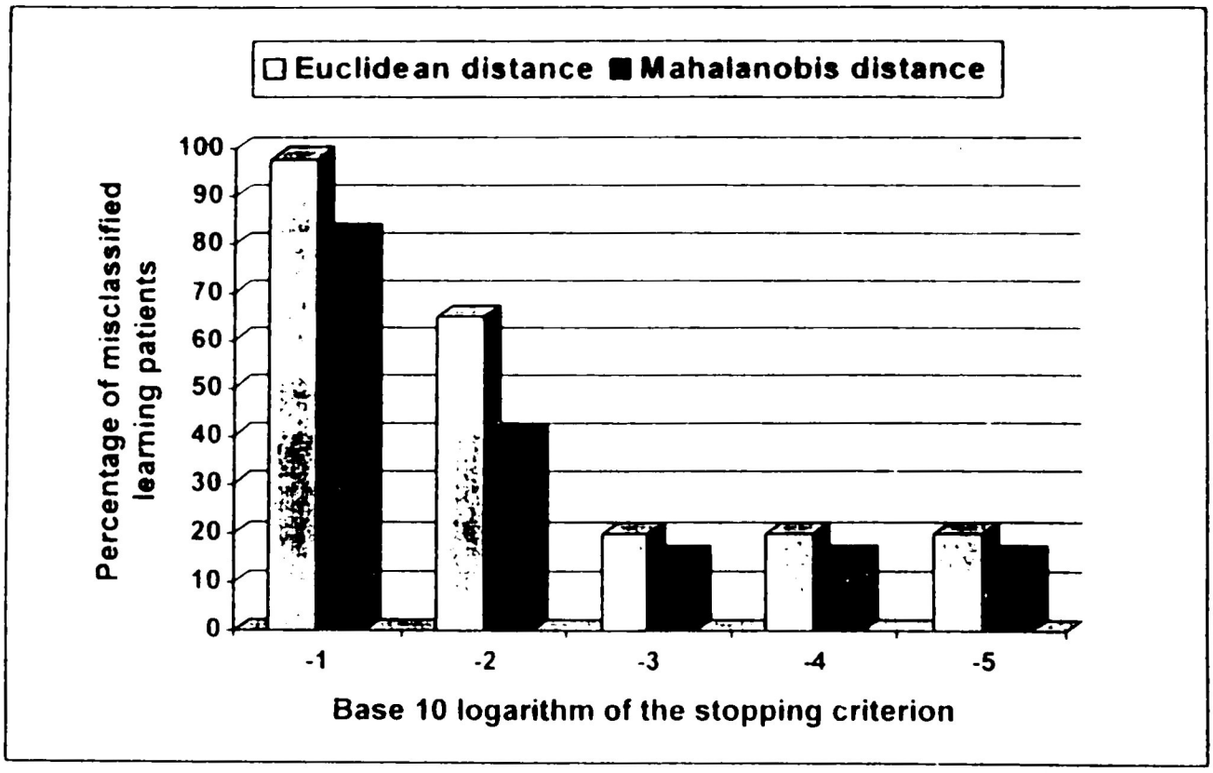

Fig. (2) : Percentage of misclassified learning patients as a function of the base 10 logarithm of the stopping criterion. Results are given by the use of both the Euclidean and the Mahalanobis distance.

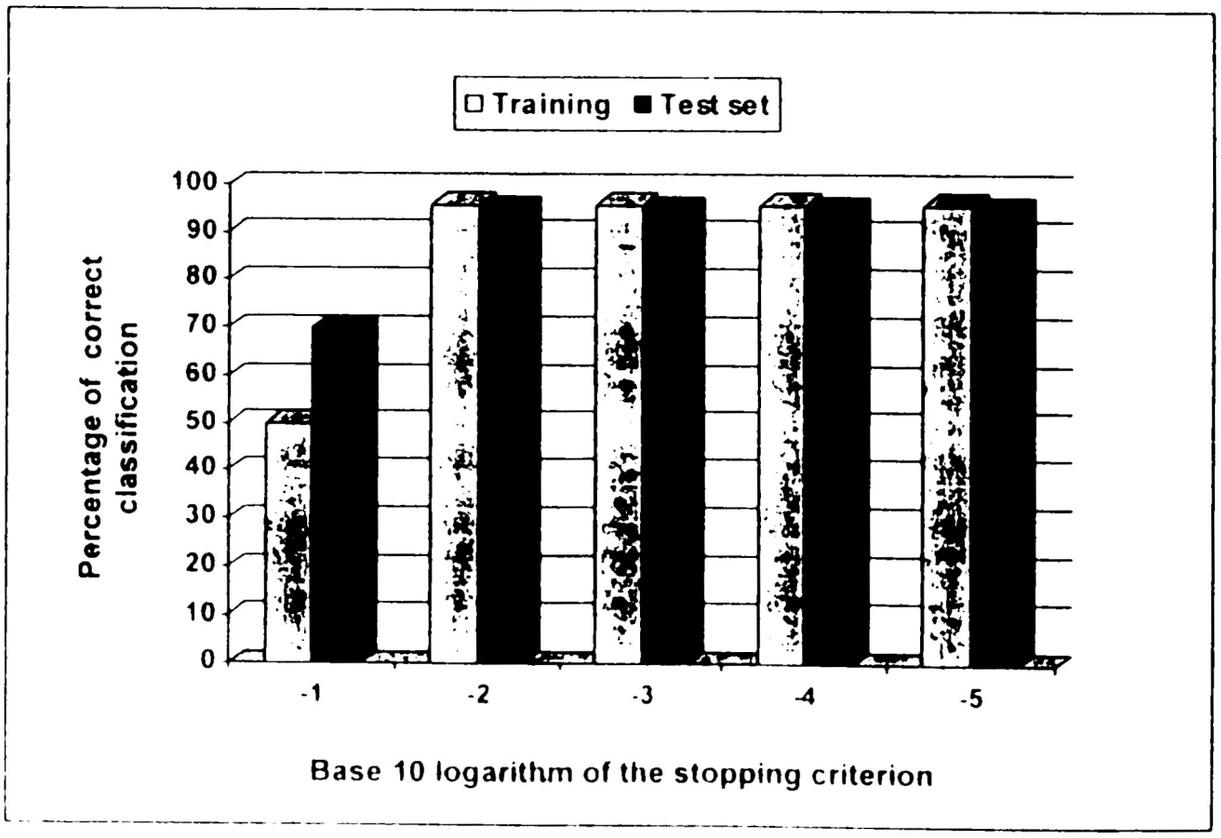

Fig. (3) : The evolution of the classification performance as a function of the base 10 logarithm of the stopping criterion.

In patient discrimination problems, the approach of achieving a classification with a reject threshold may increase the classification performances. In the improved FCMA, a way to achieve this operation was to define a reject criterion. The reject threshold represented the variance of the membership degrees of each patient.

If this variance was lower than the reject threshold, the corresponding patient was rejected. The results showed that, the variance of membership degrees is equal to 0.089. The number of rejected patients was highly sensitive even to small variations of the reject threshold. MATLAB was used in all applications.

5. Conclusion :

Fuzzy clustering, which is a non-supervised classification method, was found to be relevant for Iron deficiency Anemia identification system. The discrimination between the 3 groups of patients was obtained with good recognition rates.

The improved FCMA is not an approximation because it produces the same results as FCM but it was better than the random initialisation with respect to the computation time.

It was shown that the use of the mahalanobis distance improved the performance in comparison to the Euclidean distance. The Mahalanobis distance allowed 95.7% and 95% of the learning and test sets, while the Euclidean distance gave 94.5 % and 94% of the learning and test sets, to be correctly identified.

The Mahalanobis distance appeared to be more powerful than the Euclidean distance because the Mahalanobis distance can establish

curved as well as linear decision boundaries. Since only a limited assumption is therefore made about the shapes of the clusters, it is possible to achieve a better representation of the data, while the Euclidean distance measure is only able to create a circular bounding around the mean point of each cluster, it also allowed an improvement in the correct recognition rates. The results showed that the value of the stopping criterion ε had little influence on the recognition rates, but it had a large influence on the computation time of improved FCMA.

The number of rejected patients was highly dependent on the reject threshold parameter.

Acknowledgement :

The greatest debt and gratitude to Professor Dr. Yasser Z., (Faculty of Engineering - Cairo University) for helpful discussion on FCM algorithms and to the staff members of Machine Design Computer Lab, and to Professor Dr. Mohamed H. (El-Kasr El-Aini Hospital), Dr. Dina A. (Mataria Teaching Hospital), for providing valuable comments on IDA.

References :

Alsberg, B. and Esbensen, K. (1992), Chemom, Intell-Lab. Syst., 16, 127.

Alsberg, B. (1994), "Fast, Fuzzy C-Means Clustering of data sets with many features", Journal of computational chemistry, Vol, 16, No. 4

Bezdek, J.C. (1981), "Pattern Recognitions with Fuzzy Objective Function Algorithms", Plenum, New York.

Bezdek, J.C., Hathaway, R., Sabin, M., Tucker, W. (1987), "Convergence theory for fuzzy C-means : Counter examples and repairs", IEEE Trans. Syst. Man Cybern. 17 (5), 873-877.

Chtioui, Y., Bertrand, D., Dattee, Y., Devaux, M.F. (1996), "Identification of seeds by colour imaging. Comparison of discriminant analysis and artificial neural network", J. Food Sci. Agric., 71, 433-441.

Chtioui, Y., Dominique, B.E., Dominique, B.A., Yvette, D. (1997), "Application of Fuzzy C-Means clustering for seed discrimination", Chemometrics and intelligent laboratory system 38, 75-87.

Cobankara, V., Ozatil, D. (2001), "Cytokines and adhesive molecules in reactive thrombopoiesis" Clini Appl. Thromb. Hemost., 7 (2) : 126-30.

Dina, A., Taha, Y., Ismail, N. and Nasef, A. (2003), "Role of Thrombopoietic Cytokines in Reactive Thrombocytosis associated with Iron Deficiency Anemia". The medical Journal of Teaching Hospitals and Institutes.

Dunn, J. (1974), "A fuzzy relative ISODATA process and its detecting compact, well-separated clusters", Cybern., Vol. 3, pp. 32-57.

Ernest, B., Marshall, A., Lichtman, A., Thomas, J., Kipps, O. and William, J. (2003), Iron deficiency, Mc Graw-Hill (New York, Chicago, London).

Hamdi, A., Nilufer, G., Ismet, A. (2000), "Thrombopioetic Cytokines in patients with iron deficiency anemia with or without thrombocytosis", Acta Hematol, 103 : 152-156.

Hoffbrand, A.V., Pettit, J.E. and Moss, P.A.H. (2001), Hypochromic anemia and iron overload Blackwell Science, Oxford.

John, Y. and Reza, L. (1991), Fuzzy logic, intelligence, control and information, Prentice Hall. Upper Saddle River, New Jersey.

Kohonen, T. (1989), Self - Organisation and Associative Memory, Third ed., Spinger - Verlag.

Spath, H. (1980), Cluster Analysis Algorithms, Ellis Horwood, Chichester.

Susumu, I., Martin, J. and Robert, K. (2002), "Thrombocytosis" Medicine Journal V. 3 No. 1.