Fuzzy Sets and Systems 49 (1992) 181–191
North-Holland

On the mathematical and numerical properties of the fuzzy c-means algorithm

Shokri Z. Selim* and M.S. Kamel
Department of Systems Design, University of Waterloo, Ontario, Canada

Received October 1990
Revised January 1991

Abstract: The 'fuzzy clustering' problem is investigated. Interesting properties of the points generated in the course of applying the fuzzy c-means algorithm are revealed using the concept of reduced objective function. We investigate seven quantities that could be used for stopping the algorithm and prove relationships among them. Finally, we empirically show that these quantities converge linearly.

Keywords: Fuzzy c-means algorithm; fuzzy clustering; convergence of fuzzy c-means algorithm; stopping criteria for fuzzy c-means.

1. Introduction

In partitioning, grouping, a set of data points in the Euclidean space into a given number of clusters, if each point is restricted to belong to exactly one cluster a 'hard clustering problem' is on hand versus a 'fuzzy clustering problem' where each data point belongs to all custers with some degree of membership. Historically the latter problem evolved from the first. Fuzzy clustering should be useful in applications where clusters touch or overlap. The use of fuzzy sets in clustering goes back to the early work of Bellman et al. [1], Ruspini [14] and Gitman and Levine [10]. Dunn [8] defined the first generalization of the conventional minimum-variance hard clustering. Bezdek [2] generalized Dunn's work into a family of fuzzy clustering problems, developed an algorithm to solve the problem known as the Fuzzy C-Means Algorithm (FCMA) and gave a comprehensive treatment of the problem in [4, 3]. According to [8, 2] fuzzy clustering is achieved by solving a constrained nonlinear optimization problem (stated in Section 2).

In the current paper we shed some light on some properties of FCMA that appear to be useful for developing solution methods for the fuzzy clustering problem. We also examine the relationship among some quantities that could be used in the algorithm as well as their convergence.

In Section 2 we define the clustering problem and state the FCM algorithm. Some new properties of the problem are discussed in Section 3 using the concept of reduced objective function. In Section 4 we establish relationships among seven quantities that may be used in the stopping criterion of FCMA and report on a comprehensive empirical study of these quantities. Section 5 presents the computational experience and discusses the results on four data sets. In Section 6 we summarize the conclusions of the paper.

2. Fuzzy c-means clustering

Let $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbf{R}^s$ be a finite set of data points and let c be an integer, $1 < c < n$. Bezdek [2] proposed to solve the following mathematical program in order to have fuzzy clustering of the data into c clusters.

Problem CL.

$$\text{minimize} \quad J_m(W, Z) = \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^m d_{ij}^2 \tag{1}$$

subject to

$$\sum_{j=1}^{c} w_{ij} = 1, \quad 1 \leqslant i \leqslant n, \tag{2}$$

$$w_{ij} \geqslant 0, \quad 1 \leqslant i \leqslant n, \ 1 \leqslant j \leqslant c, \tag{3}$$

where

m is a scalar, $m > 1$,

Correspondence to: Professor M. Kamel, Dept. of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

* Dr. Selim was on sabbatical leave at the University of Waterloo. He is now back at King Fahd University of Petroleum and Minerals, Saudi Arabia.

$z_j \in \mathbf{R}^s$, $1 \leq j \leq c$, are (unknown) cluster centers,
d_{ij} is the Euclidean distance between point $\boldsymbol{x}_i$ and center z_j,
w_{ij} is the grade of association of pattern i with cluster j,
$W = \{w_{ij}\}$ is an $n \times c$ matrix, and
$Z = [z_1, z_2, \ldots, z_c]$ is an $s \times c$ matrix.

Problem CL has local minimum points which may not be global. Bezdek [2] proposed to solve the above problem by considering its first order optimality conditions which yield the following set of coupled equations:

$$z_j = \sum_{i=1}^{n} w_{ij}^m \boldsymbol{x}_i \Big/ \sum_{i=1}^{n} w_{ij}^m, \quad \forall j, \tag{4}$$

$$w_{ij} = 1 \Big/ \sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{2/(m-1)}, \quad \text{for } d_{ik} > 0, \quad \forall i, j, \tag{5}$$

$$\text{if } d_{ik} = 0 \text{ then } w_{ik} = 1 \text{ and } w_{ij} = 0 \text{ for } j \neq k. \tag{6}$$

FCMA is based on (4), (5) and (6) as given below.

Algorithm FCM.
Initialization. Select membership functions $W^{(1)}$ arbitrarily. Set $k = 1$.
Step 1. Compute $Z^{(k)}$ using $W^{(k)}$ and (4).
Step 2. Compute $W^{(k+1)}$ using $Z^{(k)}$, (5) and (6).
Step 3. If $f(W^{(k)}, W^{(k+1)}) < \epsilon$ stop, where $\epsilon > 0$ is a small scalar and f is some function. Otherwise set $k = k + 1$ and go to Step 1.

The algorithm could start by initializing $Z^{(1)}$ rather than $W^{(1)}$. The algorithm has been shown in [5] to converge to points satisfying the following conditions:

$$J_m(W^*, Z^*) \leq J_m(W, Z^*), \quad \forall W \in \mathbf{M}_{nc},$$

$$J_m(W^*, Z^*) \leq J_m(W^*, Z), \quad \forall Z \subset \mathbf{R}^s,$$

if all $w_{ij} > 0$, otherwise the first condition becomes strict inequality, where $\mathbf{M}_{nc}$ is the set of all $n \times c$ real matrices satisfying (2) and (3) and the cluster non-degeneracy condition:

$$\sum_{i=1}^{n} w_{ij} > 0, \quad 1 \leq j \leq c.$$

Points satisfying the above conditions may not be local minima of Problem CL. See [5, 18, 11, 21, 12] for the characterization of the terminal points produced by the algorithm.

3. Properties of the problem

To facilitate the understanding of the problem we use the concept of reduced objective function, originally introduced in [15] which proved to be very useful for studying the properties of solutions produced by FCMA [18, 11].

Definition. The reduced objective function of J_m is given by

$$\psi_m(W) = \min_Z J_m(W, Z),$$

where $W = \{w_{ij}\}$, $w_{ij} \geq 0$.

The following problem is equivalent to Problem CL.

Problem EP.
minimize $\psi_m(W)$
subject to

$$\sum_{j=1}^{c} w_{ij} = 1,$$

$$w_{ij} \geq 0, \quad 1 \leq i \leq n, \ 1 \leq j \leq c.$$

We note here that ψ_m is nonconvex and may have local minimum and saddle points. One could utilize the properties of Problem EP in developing algorithms to solve Problem CL. In [18, 11, 15], ψ_m is used to derive local optimality conditions of the termination points of FCMA. The gradient and Hessian of ψ_m are also given in [18].

From (2) we have $w_{ir} = 1 - \sum_{j=1, j\neq r}^{c} w_{ij}$, for some i and r. Substituting for w_{ir} in ψ_m one obtains $\Psi_m(W^r)$, where W^r is an $n \times (c-1)$ real matrix which contains the columns of W except for the r-th column. The following problem is then equivalent to Problems CL and EP:

minimize $\Psi_m(W^r)$
subject to

$$w_{ij} \geq 0, \quad 1 \leq i \leq n, \ 1 \leq j \leq c, j \neq r, \tag{7}$$

$$\sum_{j=1, j\neq r}^{c} w_{ij} \leq 1. \tag{8}$$

In Theorem 3.1 we prove an interesting property of the stationary points of Ψ_m.

Lemma 3.1.

$$\partial \Psi_m / \partial w_{pq} = m w_{pq}^{m-1} \|x_p - z_q^*\|^2 - m w_{pr}^{m-1} \|x_p - z_r^*\|^2, \tag{9}$$

where $q \neq r$, Z^* is the minimum of $J_m(W, Z)$ for a fixed W.

Proof.

$$\Psi_m(W^r) = \sum_{i=1}^{n} \sum_{j=1, j\neq r}^{c} w_{ij}^m \|x_i - z_j^*\|^2 + \sum_{i=1}^{n} \Big(1 - \sum_{j=1,j\neq r}^{c} w_{ij}\Big)^m \|x_i - z_r^*\|^2.$$

Hence,

$$\partial \Psi_m / \partial w_{pq} = m w_{pq}^{m-1} \|x_p - z_q^*\|^2 - 2(\partial z_j / \partial w_{pq})' \sum_{i=1}^{n} w_{iq}^m (x_i - z_q^*) - m\Big(1 - \sum_{j=1,j\neq r}^{c} w_{pj}\Big)^{m-1} \|x_p - z_r^*\|^2.$$

From (4), the sum in the middle term vanishes and the above expression simplifies to (9).

Theorem 3.1. *If m is not an odd number then the stationary points of Ψ_m satisfy* (7) *and* (8).

Proof. Equation (9) must vanish at the stationary points of Ψ_m yielding the following system of equations:

$$w_{ij}^{m-1} d_{ij}^2 = w_{ir}^{m-1} d_{ir}^2, \quad 1 \leqslant i \leqslant n, \ 1 \leqslant j \leqslant c. \tag{10}$$

If m is not integer then w_{ij}^m is not defined for negative w_{ij} and hence a real solution satisfying (10) will saitsfy (3). Furthermore if m is even and if $w_{pq} < 0$ for some p and q, then, from (10) this implies that $w_{iq} < 0$ for all j, which in turn violates (2). The latter condition is already implied in Ψ_m and hence $w_{ij} > 0$ for all i and j. Finally if (8) is violated then $w_{ir} < 0$ and hence $w_{iq} < 0$ for all j, which is a contradiction. Therefore, (8) is also satisfied. This completes the proof.

The above theorem asserts that if m is not an odd number then one could resort to unconstrained optimization algorithms to find the minimum of Ψ_m disregarding the constraints. The solution according to the theorem will satisfy the constraints.

Let

$$Y_i = w_{ij}^{m-1} d_{ij}^2, \quad 1 \leqslant i \leqslant n.$$

Then

$$w_{ij} = d_{ij}^{2/(1-m)} Y_i^{1/(m-1)}, \quad \forall i, j \tag{11}$$

and by invoking (2) one obtains

$$Y_i = \Big[\sum_{j=1}^{c} d_{ij}^{2/(1-m)}\Big]^{1-m} \tag{12}$$

Equations (11) and (12) do not lead to a new algorithm for solving the fuzzy clustering problem. The formula used in FCMA could be simplified to (11) and (12) and could be used to enhance computational efficiency. They also lead to an interesting result with respect to the objective function.

Theorem 3.2.

$$J_m(W, Z) = \sum_{i=1}^{n} Y_i, \tag{13}$$

$$\partial \psi_m(W) / \partial w_{ij} = m Y_i. \tag{14}$$

Proof. The objective function as defined in (1) can be rewritten as

$$J_m(W, Z) = \sum_{i=1}^{n} \sum_{j=1}^{c} Y_i^{m/(m-1)} d_{ij}^{2/(1-m)} = \sum_{i=1}^{n} Y_i^{m/(m-1)} \sum_{j=1}^{c} d_{ij}^{2/(1-m)} = \sum_{i=1}^{n} Y_i.$$

Equation (14) follows directly from equation (18) in [18]. This completes the proof.

The beauty of (13) is that the Y_i's are computed anyway in order to obtain the w_{ij}'s. These same values could be used to compute J_m. The original expression for J_m calls for computing w_{ij}^m; an expensive function, to be repeated nc times, nc multiplications and nc additions, while (13) calls for only n additions and n exponentiations. Equation (11) leads to the following interesting results:

Theorem 3.3. *Let $W_m^*(Z)$ be the optimal solution of $J_m(W, Z)$ for Z fixed and W satisfying* (2) *and* (3). *Furthermore let W_m^* and Z_m^* be the optimal*

solution of Problem CL. *Then*:
(i) $cJ_m(W^*_m(Z), Z) = J_{m-1}(W^*_m(Z), Z)$,
(ii) $cJ_m(W^*_m, Z^*_m) \geqslant J_{m-1}(W^*_{m-1}, Z^*_{m-1})$,
(iii) $c^r J_{m+r}(W^*_{m+r}, Z^*_{m+r}) \geqslant J_m(W^*_m, Z^*_m)$.

Proof. Let $W^*_m(Z) = \{w^*_{ij}\}$. Then

$$J_{m-1}(W^*_m(Z), Z) = \sum_{i=1}^{n}\sum_{j=1}^{c} w_{ij}^{*m-1} d_{ij}^2 = c\sum_{i=1}^{n} Y_i = cJ_m(W^*_m(Z), Z).$$

To show (ii) note that

$$J_{m-1}(W^*_m(Z), Z) \geqslant J_{m-1}(W^*_{m-1}, Z^*_m).$$

Finally, (iii) could be shown by is applying (ii) r times.

Part (ii) of Theorem 3.3 could be used to establish a lower bound for the objective function for some m given that at $m-1$. For example if $m=2$ then the right hand side of (ii) will correspond to the objective function value associated with the hard clustering problem. Currently there are algorithms for obtaining the global solution of the latter problem [16, 17]. If one is interested in obtaining the global solution of the fuzzy problem then a possible approach is to apply FCMA several times starting each time with a different initialization. Among the several solutions obtained the one yielding the least J_m is selected. This process could be stopped whenever a solution is obtained which is within acceptable range from the lower bound generated by (ii). The bound given in (iii) could be used similarly. Equations (12) and (13) will be further used to show some results in Section 4.

4. Stopping criteria and convergence of FCMA

In this section we discuss the behaviour of some stopping criteria that could be used with FCMA. We restrict the discussion to those criteria where the following quantities are computed:

$$Q_v(k) = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |w_{ij}(k) - w_{ij}(k-1)|^v\right)^{1/v},$$

$v = 1, 2$ and ∞,

$$R_v(k) = \left(\sum_{j=1}^{c}\sum_{l=1}^{s} |z_{jl}(k) - z_{jl}(k-1)|^v\right)^{1/v},$$

$v = 1, 2$ and ∞,

$$T(k) = J_m(W^{(k-1)}, Z^{(k-1)}) - J_m(W^{(k)}, Z^{(k)}).$$

If $v=\infty$, Q_v and R_v simplify to

$$Q_\infty(k) = \max_{i,j} |w_{ij}(k) - w_{ij}(k-1)|,$$

$$R_\infty(k) = \max_{j,l} |z_{jl}(k) - z_{jl}(k-1)|.$$

In case of a matrix Q_2 and R_2 are called Frobenius norms. On the other hand if the rows of a matrix are considered to form a single vector then if $v=1$, the corresponding norm is the rectlinear or l_1-norm. If $v=2$, the norm is the Euclidean or l_2-norm. While if $v=\infty$, the norm is called Chebychev, sup, or l_∞-norm. The quantities R_v and T depend on the magnitude of X. If each pattern vector is multiplied by a scalar, say, ρ then R_v will be multiplied by ρ and T by ρ^2.

In Theorems 4.1 to 4.4 below we prove some relationships among the above quantities. In the second part of this section we report our empirical study of these quantities. Section 4.3 contains an empirical study of the convergence of these quantities.

4.1. Theoretical relationships among Q_v, R_v and T

The following theorem gives relationships among Q_v (R_v) for $v = 1, 2$, and ∞.

Theorem 4.1. (1) $cnQ_\infty \geqslant Q_1$.
(2) $\sqrt{cn}\, Q_2 \geqslant Q_1$.
(3) $Q_1 \geqslant Q_2 \geqslant Q_\infty$.
(4) $csR_\infty \geqslant R_1$.
(5) $\sqrt{cs}\, R_2 \geqslant R_1$.
(6) $R_1 \geqslant R_2 \geqslant R_\infty$.

Proof. See Stewart [20], page 170.

In the Theorem 4.2 a relationship among Q_2, R_2, and T is established. But first the following two lemmas are proven.

Lemma 4.1.

$$J(W^{(k)}, Z^{(k-1)}) - J(W^{(k)}, Z^{(k)}) \leqslant nR_2^2.$$

Proof. Consider a point x_i and a center $z_j^{(k-1)}$. Then

$$x_i - z_j^{(k-1)} = (x_i - z_j^{(k)}) + (z_j^{(k)} - z_j^{(k-1)}).$$

Squaring both sides yields

$$\begin{aligned}\|x_i - z_j^{(k-1)}\|^2 = \|x_i - z_j^{(k)}\|^2 \\ + 2(x_i - z_j^{(k)})'(z_j^{(k)} - z_j^{(k-1)}) \\ + \|z_j^{(k)} - z_j^{(k-1)}\|^2.\end{aligned}$$

Multiplying by $w_{ij}^m(k)$ and summing over all values of i and j one obtains

$$\begin{aligned}J(W^{(k)}, Z^{(k-1)}) &= J(W^{(k)}, Z^{(k)}) \\ &+ 2\sum_j (z_j^{(k)} - z_j^{(k-1)})' \sum_i w_{ij}^m(k)(x_i - z_j^{(k)}) \\ &+ \sum_j \|z_j^{(k)} - z_j^{(k-1)}\|^2 \sum_i w_{ij}^m(k).\end{aligned} \tag{15}$$

But $\sum_i w_{ij}^m(k)(x_i - z_j^{(k)}) = 0$. Hence (15) simplifies to

$$\begin{aligned}J(W^{(k)}, &Z^{(k-1)}) - J(W^{(k)}, Z^{(k)}) \\ &= \sum_j \|z_j^{(k)} - z_j^{(k-1)}\|^2 \sum_i w_{ij}^m(k) \\ &\leqslant n \sum_j \|z_j^{(k)} - z_j^{(k-1)}\|^2 \\ &= nR_2^2.\end{aligned} \tag{16}$$

This completes the proof.

Lemma 4.2. *Given x and y such that $1 \geqslant x, y \geqslant 0$, then*

$$x^m - y^m \leqslant m\,|x-y| \quad \textit{for } m \geqslant 1. \tag{17}$$

Proof. Assume that $x > y$ otherwise the proof is trivial. To show (17) we will show that

$$mx - x^m \geqslant my - y^m.$$

Note that $f(u) = mu - u^m$ is an increasing function for $0 \leqslant u \leqslant 1$ and $m > 1$. Since $x > y$ then $f(x) > f(y)$. This completes the proof.

In the following theorem a bound on T is established using the results of the above lemmas.

Theorem 4.2.

$$T \leqslant nR_2^2 + mLQ_1$$

where

$$L = \max_{i,t} \sum_l (x_{il} - x_{tl})^2.$$

Proof.

$$\begin{aligned}T = J(W^{(k-1)}, Z^{(k-1)}) - J(W^{(k)}, Z^{(k-1)}) \\ + J(W^{(k)}, Z^{(k-1)}) - J(W^{(k)}, Z^{(k)}).\end{aligned} \tag{18}$$

Now

$$\begin{aligned}J(W^{(k-1)}, &Z^{(k-1)}) - J(W^{(k)}, Z^{(k-1)}) \\ &= \sum_{i,j} (w_{ij}^m(k-1) - w_{ij}^m(k))\, \|x_i - z_j^{(k-1)}\|^2 \\ &\leqslant L \sum_{i,j} (w_{ij}^m(k-1) - w_{ij}^m(k)).\end{aligned} \tag{19}$$

Using Lemma 4.2 one obtains

$$\begin{aligned}L \sum_{i,j} &(w_{ij}^m(k-1) - w_{ij}^m(k)) \\ &\leqslant mL \sum_{i,j} \|w_{ij}(k-1) - w_{ij}(k)| \\ &= mLQ_1.\end{aligned} \tag{20}$$

Combining (16), (18), and (20) yields the result.

The results of the following lemma are used in Theorem 4.3 to develop a tighter upper limit on T.

Lemma 4.3. *Let* $L = \max_{i,t} \sum_l (x_{il} - x_{tl})^2$ *and* $D_{ij} = d_{ij}^2 = \sum_l (x_{il} - z_{jl})^2$. *Then*

(i) $Y_i \leqslant L/c^{m-1}$,
(ii) $\partial Y_i / \partial D_{ij} = w_{ij}^m$, *and*
(iii) $J_m(W, Z) \leqslant nL/c^{m-1}$.

Proof. From (12) we have

$$\begin{aligned}Y_i &= \left[\sum_{j=1}^{c} d_{ij}^{2/(1-m)}\right]^{1-m} \\ &\leqslant \left[\sum_{j=1}^{c} L^{1/(1-m)}\right]^{1-m} = L/c^{m-1}.\end{aligned} \tag{21}$$

To show (ii) note that

$$\begin{aligned}w_{ij}^m &= \partial J / \partial D_{ij} \\ &= \partial J / \partial Y_i \cdot \partial Y_i / \partial D_{ij} = \partial Y_i / \partial D_{ij}.\end{aligned}$$

Result (iii) follows directly from (i) and (13).

Theorem 4.3. $T \leqslant mLQ_1/c^{m-1}$ *for* Q_1 *sufficiently small.*

Proof. Note that

$$\psi_m(W^{(k)}) = \min_Z J_m(W^{(k)}, Z) = J_m(W^{(k)}, Z^{(k)}). \tag{22}$$

From (22),

$$T = \psi_m(W^{(k-1)}) - \psi_m(W^{(k)}).$$

The total differential of ψ is given by

$$\Delta\psi \simeq \sum_{i,j} \partial\psi/\partial w_{ij}\, \Delta w_{ij} \tag{23}$$

where $\Delta w_{ij} = |w_{ij}(k-1) - w_{ij}(k)|$ is small and

$$\Delta\psi = \psi_m(W^{(k-1)}) - \psi_m(W^{(k)}).$$

From [18] we have

$$\partial\psi/\partial w_{ij} = m w_{ij}^{m-1} d_{ij}^2. \tag{24}$$

Substituting from (24) into (23) one obtains

$$T \simeq \sum_{i,j} m w_{ij}^{m-1} d_{ij}^2\, \Delta w_{ij} = \sum_{i,j} m Y_i\, \Delta w_{ij}. \tag{25}$$

Substituting (21) into (25) we get

$$\begin{aligned} T &\leqslant (mL/c^{m-1}) \sum_{i,j} \Delta w_{ij} \\ &= (mL/c^{m-1}) \sum_{i,j} |w_{ij}(k-1) - w_{ij}(k)| \\ &= mLQ_1/c^{m-1} \end{aligned} \tag{26}$$

This completes the proof.

In the next theorem a relation between T and R_1 is established.

Theorem 4.4. *Let $M = \max_{i,t,l} |x_{i,l} - x_{t,l}|$. Then*

$$T \leqslant 2nMR_1 \tag{27}$$

for R_1 sufficiently small.

Proof. Let

$$D_{i,j} = d_{i,j}^2 = \sum_{l=1}^{s} (x_{il} - z_{jl})^2.$$

Furthermore let $\mathrm{d}J$, $\mathrm{d}Y_i$, $\mathrm{d}D_{ij}$ and $\mathrm{d}z_{jl}$ be the differentials of J, Y_i, D_{ij} and z_{jl} respectively. Then from (13),

$$\begin{aligned} \mathrm{d}J &= \sum_i (\partial J/\partial Y_i)\, \mathrm{d}Y_i = \sum_i \mathrm{d}Y_i \\ &= \sum_i \sum_j (\partial Y_i/\partial D_{ij})\, \mathrm{d}D_{ij} \\ &= \sum_i \sum_j w_{ij}^m\, \mathrm{d}D_{ij} \quad \text{(from Lemma 4.3(ii))} \\ &\leqslant \sum_i \sum_j \sum_l (\partial D_{ij}/\partial z_{jl})\, \mathrm{d}z_{jl} \\ &= 2 \sum_i \sum_j \sum_l (z_{jl} - x_{il})\, \mathrm{d}z_{jl} \\ &\leqslant 2M \sum_i \sum_j \sum_l \mathrm{d}z_{jl}. \end{aligned}$$

Let ΔJ and Δz_{jl} be the total differentials of J and z_{jl} respectively.

If $\Delta z_{jl} = |z_{jl}(k) - z_{jl}(k-1)|$ is sufficiently small then $\Delta J \cong \mathrm{d}J$ and $\Delta z_{jl} \simeq \mathrm{d}z_{jl}$. Hence

$$\begin{aligned} \Delta J &\leqslant 2M \sum_i \sum_j \sum_l |z_{jl}(k) - z_{jl}(k-1)| \\ &= 2nMR_1. \end{aligned}$$

This completes the proof.

It is straight forward to show that scaling of the data points will not affect the inequalities developed in the theorems and lemmas of this section.

4.2. *Empirical study of the relationships among Q_v, R_v and T*

We have conducted a comprehensive study of the above quantities using four published data sets namely, the British towns data [13] (the first 50 samples with the first four principal components), the Fossil data [6], the German towns data [19], and the Iris data [9]. For each data set forty initial clusterings of the data were randomly generated and the FCMA was run starting with each. The experiments were conducted using a μ vax II computer running Unix operating system. All calculations were in double precision. The following observations have been made:

Observation 1. The bounds on T provided in Theorems 4.3 and 4.4 are valid for Q_1 and R_1 small, which is the case at the final iterations of the algorithm, while the bound provided in Theorem 4.2 is valid at any iteration. Interestingly enough our experiments showed that the bounds of Theorems 4.3 and 4.4 are satisfied at any iteration of the algorithm.

Observation 2. The ratio of the right hand side to the left hand side of each of the inequalities given in Theorems 4.2 to 4.4 increases as the number of iterations increases.

Table 1. The ratios of the bounds of Theorems 4.2 to 4.4 to T

	$(nR_2^2 + mLQ_1)/T$		$mLQ_1/(c^{m-1}T)$		$2nMR_1/T$	
Data set	min	max	min	max	min	max
British towns	140	∞	40	∞	125	∞
Fossil data	75	10^{10}	25	$10^{10}/3$	75	10^{10}
German towns	100	∞	35	10^{10}	45	10^{10}
Iris data	200	∞	65	∞	100	∞

Table 1 shows the maximum and minimum values these ratios achieve for each of the data sets considered. The value of ∞ shown in the table corresponds to $T = 0.0$.

Observation 3. T could become negative, i.e. J_m has increased at some iteration which contradicts proven theory [5]. This indicates loss of precision due to computational rounding errors.

Observation 4. As the number of iterations increases the order of magnitude of T becomes much less than that of any other quantity. To illustrate this behaviour let k_ϵ denote the first iteration at which no quantity exceeds ϵ and let k_e be the last iteration at which T is non-negative, i.e.

$$k_\epsilon = \min\{k\colon R_v(k),\ Q_v(k),\ T(k) \leq \epsilon\}$$

and

$$k_e = \max\{k\colon T(k) \geq 0\}.$$

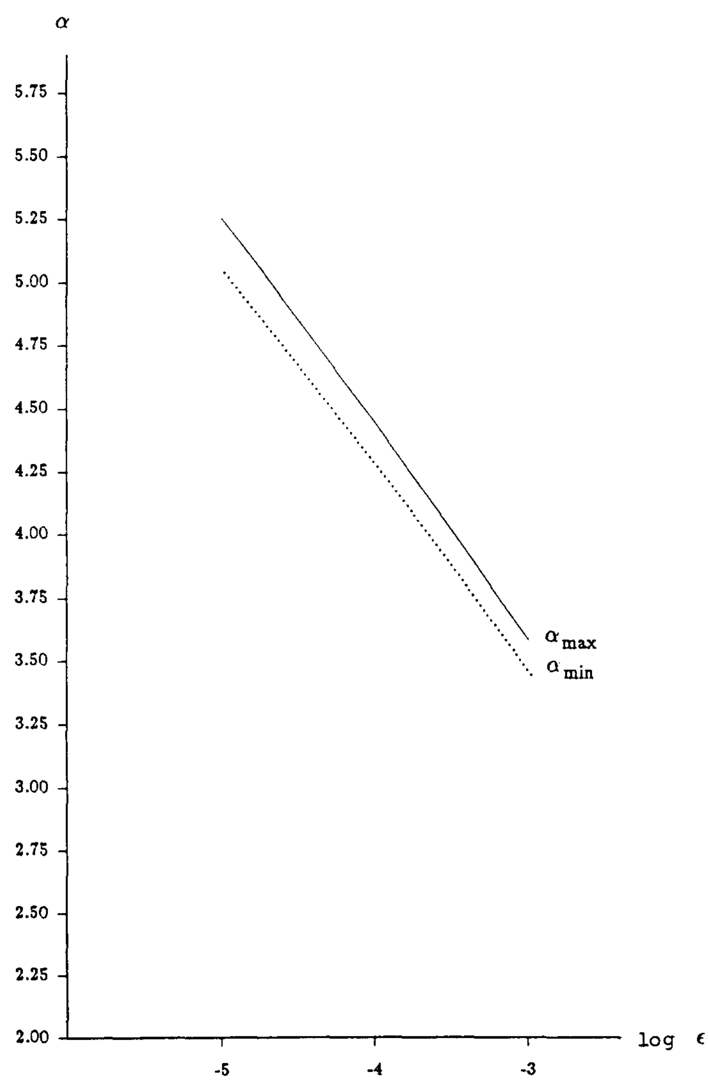

Fig. 1. $\alpha_{\max}$ and $\alpha_{\min}$ for the British towns data.

As FCMA is being run the ratios $T(k)/R_v(k)$ and $T(k)/Q_v(k)$ are computed for $v = 1, 2, \infty$. We make the remark here that the reciprocal of the above ratios was not considered because $T(k)$ could vanish. Since the above ratios decrease with the iterations the average of each ratio is computed for a particular range of iterations as given by the following:

$$A(v, \epsilon) = \sum_{k=k_\epsilon}^{k_e} (T(k)/R_v(k))/(k_e - k_\epsilon),$$

$$B(v, \epsilon) = \sum_{k=k_\epsilon}^{k_e} (T(k)/Q_v(k))/(k_e - k_\epsilon),$$

$v = 1, 2$, and ∞, $\epsilon = 10^{-15}, 10^{-4}$, and 10^{-3}.

The logarithm to base 10 of the reciprocal of any of the above averages is the difference in order of magnitude of R or Q and T. Those differences become larger as ϵ becomes smaller. The above calculations were performed on each data set forty times each time starting from a different clustering of the data points. The maximum and minimum difference in magnitude over the forty runs was recorded.

Let

$$\alpha_{max}(v, \epsilon) = \max \log_{10} 1/A(v, \epsilon),$$

$$\alpha_{min}(v, \epsilon) = \min \log_{10} (1/A(v, \epsilon),$$

$$\gamma_{max}(v, \epsilon) = \max \log_{10} 1/B(v, \epsilon)$$

and

$$\gamma_{min}(v, \epsilon) = \min \log_{10} 1/B(v, \epsilon).$$

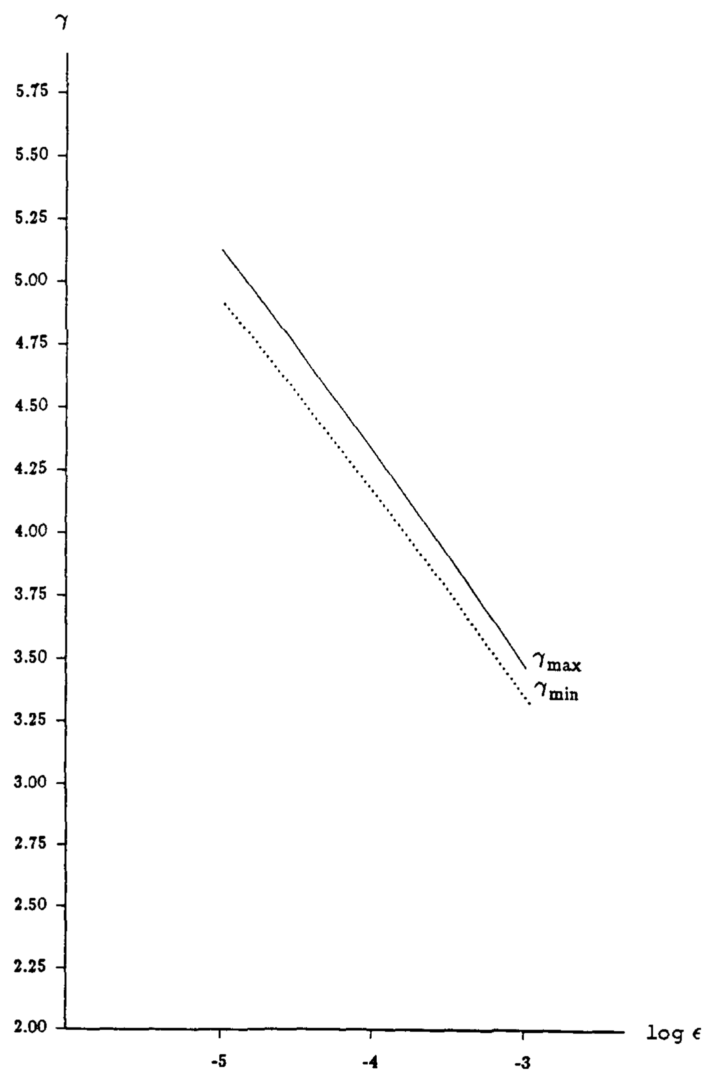

Fig. 2. γ_{max} and γ_{min} for the British towns data.

In the graphs of Figures 1 and 2 we consider the case $v = \infty$ since, $R_\infty \leqslant R_v$ and $Q_\infty \leqslant Q_v$, for all v. The graphs corresponding to any other value of v will be even higher than the ones shown.

In the case of the British towns data, Figure 1 shows the behaviour of $\alpha_{\max}(\infty, \epsilon)$ and $\alpha_{\min}(\infty, \epsilon)$ versus $\log_{10} \epsilon$, while Figure 2 shows $\gamma_{\max}(\infty, \epsilon)$ and $\gamma_{\min}(\infty, \epsilon)$.

The graphs clearly show that the difference in the order of magnitudes of any norm and T is large and it gets larger as the iterations increase. It should be mentioned that A and B depend on the scale of the data. If each data point is multiplied by ρ then $\alpha_{\min}$ and $\alpha_{\max}$ will have the term $-\log_{10} \rho$ added to each. On the other hand $\gamma_{\min}$ and $\gamma_{\max}$ will have the term $-2 \log_{10} \rho$ added to each.

4.3. Empirical study of the order of convergence

We end this section with an empirical study on the order of convergence of FCMA. Recall that a sequence $g^{(k)}$ converges to g with order p and asymptotic error constant β if the following is true [7]:

$$\lim_{k \to \infty} \|g^{(k)} - g\| / \|g^{(k-1)} - g\|^p = \beta. \tag{28}$$

We consider the cases $g^{(k)} = J(k)$, $Z^{(k)}$ and $W^{(k)}$ and the norms corresponding to $v = 1, 2,$ and ∞. Furthermore, we assume $g = J(k_e)$, $Z^{(k_e)}$

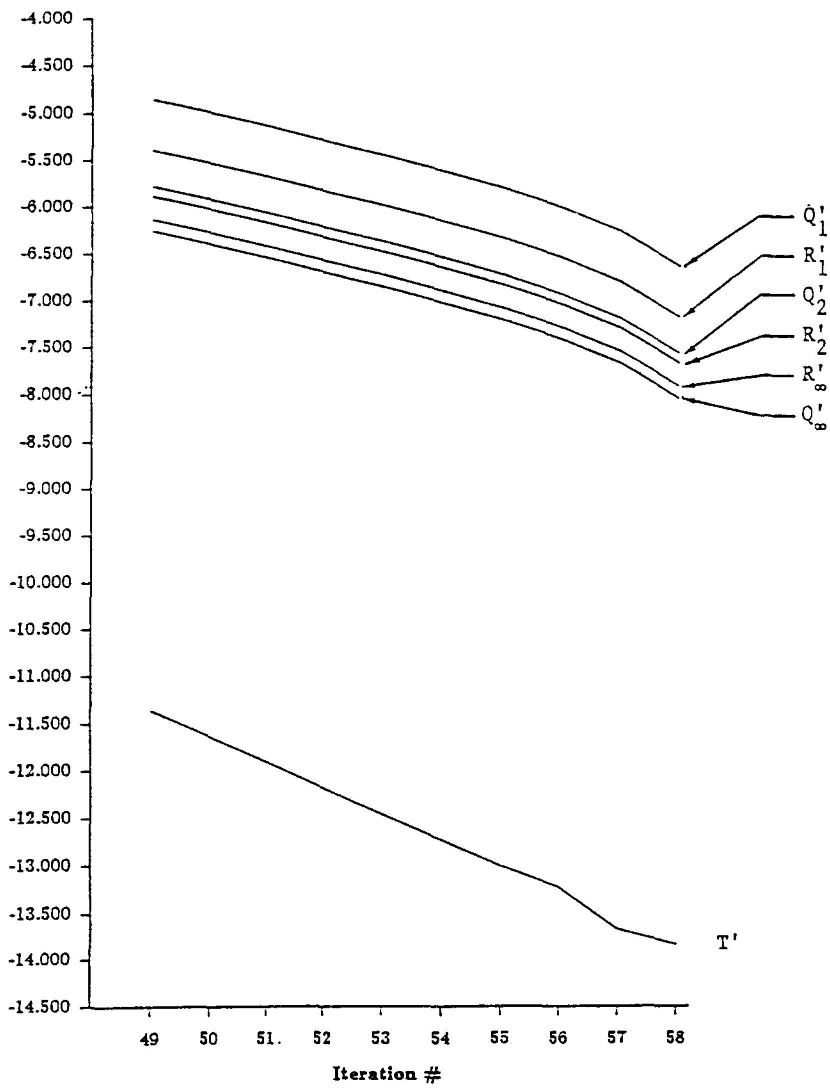

Fig. 3. Convergence behaviour of T', Q', and R'.

and $W^{(k_e)}$. So $\log \|g^{(k)} - g\|$ will have the following three forms depending on which variable is being considered:

$$Q'_v(k) = \log\Big(\sum_{i=1}^{n}\sum_{j=1}^{n} |w_{ij}(k) - w_{ij}(k_e)|^v\Big)^{1/v},$$

$v = 1, 2$ and ∞,

$$R'_v(k) = \log\Big(\sum_{j=1}^{c}\sum_{l=1}^{s} |z_{jl}(k) - z_{jl}(k_e)|^v\Big)^{1/v},$$

$v = 1, 2$ and ∞,

$$T'(k) = \log |J_m(W^{(k)}, Z^{(k)}) - J_m(W^{(k_e)}, Z^{(k_e)})|.$$

Consider the equation

$$\|g^{(k)} - g\| / \|g^{(k-1)} - g\|^{p_k} = \beta_k. \tag{29}$$

As k approaches k_e, p_k and β_k approach p and β of (28) respectively. Taking the logarithm of both sides of (29) yields

$$\log \beta_k = -(p_k \log \|g^{(k-1)} - g\| - \log \|g^{(k)} - g\|). \tag{30}$$

In Figure 3 we plot the quantities $T'(k)$, $Q'_v(k)$ and $R'_v(k)$, $k_e - 10 \leqslant k \leqslant k_e$, for a run of the British towns data.

Each curve in the figure shows linear behaviour for most of the iterations indicating that the order of convergence of the respective sequence is linear. From (30), β_k becomes inversely proportional to the change in $\log \|g^{(k)} - g\|$ from one iteration to the next and hence β_k is inversely proportional to the slope of the graph since $p = 1$. From Figure 3 it is clear that the curve corresponding to T' is steeper than any of the others and hence the corresponding β is smaller than any of the others.

In conclusion, the order of convergence of the sequences under consideration is the same, namely, linear while the asymptotic error constant is the same for all sequences except for T' which is also the smallest. This conclusion applies to all data sets studied.

5. Computational experience

An implementation of the FCMA using (11) has been tested on the data sets introduced in Section 4.2. Its performance in terms of the number of iterations and execution time has been compared for different stopping criteria. The values are then averaged over forty runs corresponding to forty different initial clustering of the data. The experiments were repeated for several values of m and c. Representative results comparing the performance of the algorithm using R_1 as a stopping criterion versus that using T are shown in Tables 2 to 5. These results show that the use of T reduces the number of iterations and required CPU time. The results are also consistent over all the runs. It should be noted that starting from the same initial

Table 2. Results of the British towns data, for $c = 4$

m	Criterion	Mean no. of iterations	CPU time in seconds
1.5	R_1	28.8	9.2
	T	19.6	6.2
2.0	R_1	33.6	10.8
	T	20.3	6.6
2.5	R_1	34.8	11.2
	T	20.2	6.4
3.0	R_1	39.1	12.6
	T	21.1	6.7

Table 3. Results of the Fossil data, for $c = 3$

m	Criterion	Mean no. of iterations	CPU time in seconds
1.5	R_1	21.0	9.8
	T	15.4	7.2
2.0	R_1	27.8	13.0
	T	19.8	9.2
2.5	R_1	54.1	25.2
	T	34.0	15.5
3.0	R_1	339.1	158.2
	T	69.0	31.4

Table 4. Results of the German towns data, for $c = 3$

m	Criterion	Mean no. of iterations	CPU time in seconds
1.5	R_1	40.9	11.1
	T	32.0	8.6
2.0	R_1	41.1	11.1
	T	32.6	8.7
2.5	R_1	54.5	15.6
	T	45.3	12.2
3.0	R_1	102.3	27.7
	T	83.1	22.4

Table 5. Results of the Iris data, for $c = 3$

m	Criterion	Mean no. of iterations	CPU time in seconds
1.5	R_1	17.6	13.1
	T	12.4	9.3
2.0	R_1	20.7	15.4
	T	14.7	10.9
2.5	R_1	20.9	15.6
	T	14.4	10.6
3.0	R_1	22.7	16.9
	T	15.1	11.1

clustering the value of J_m obtained using the different stopping criteria is the same.

6. Conclusions

In this paper, we have investigated the FCMA. The concept of reduced objective function was used to reveal several new properties of points generated by the algorithm. One of the properties has the potential of leading to new algorithms for solving Problem CL. Another property includes a lower bound on the objective function which has a potential of developing criteria to be used in search for global solutions.

We also studied some quantities that could be used to stop FCMA. Relationships among these were developed. These relationships should assist the user of FCMA in selecting a stopping criterion. Finally, an empirical study concluded, for few data sets, that all the quantities considered converge linearly.

The properties also show new relationships with respect to the objective function being minimized and facilitates the use of a criterion based on testing the reduction in the objective function value.

Acknowledgement

This work was partially supported by the Natural Science and Engineering Research Council of Canada through a research grant to the second author.

Support from King Fahd University of Petroleum and Minerals through sabbatical leave for the first author is also acknowledged.

The computational results of Section 5 were obtained by Mr. Heru Natalisa.

References

[1] R.E. Bellman, R. Kalaba and L.A. Zadeh, Abstraction and pattern classification, *J. Math. Anal. Appl.* **13** (1966) 1–7.

[2] J.C. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. Dissertation, Appl. Math., Cornell Univ., Ithaca, NY (1973).

[3] J.C. Bezdek, A convergence theorem for the Fuzzy ISODATA clustering algorithms, *IEEE Trans. Pattern Anal. Machine Intelligence* **2**(1) (1980) 1–8.

[4] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum, New York, 1981).

[5] J.C. Bezdek, R. Hathaway, M. Sabin and W. Tucker, Convergence theory for fuzzy c-Means: Counter examples and repairs, *IEEE Trans. System Man Cybernet.* **17**(5) (1987) 873–877.

[6] H. Chernoff, The use of faces to represent points in k-dimensional space graphically, *J. Amer. Statist. Assoc.* **68** (1973) 361–368.

[7] G. Dahlquist and A. Björck, *Numerical Methods* (Prentice-Hall, Englewood Cliffs, NJ, 1979).

[8] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. Cybernet.* **3** (1973) 32–57.

[9] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. of Eugenics* **3** (1936) 179–188.

[10] I. Gitman and M.D. Levine, An algorithm for detecting unimodal fuzzy sets and its application as a clustering technique, *IEEE Trans. Comput.* **19** (1970) 583–593.

[11] M.A. Ismail and S.Z. Selim, Fuzzy c-means: Optimality of solutions and effective termination of the algorithm, *Pattern Recognition* **19**(6) (1986) 481–485.

[12] T. Kim, J.C. Bezdek, R. Hathaway, Optimality test for fixed points of the fuzzy c-means algorithm, *Pattern Recognition* **21**(6) (1988) 651–663.

[13] C.A. Moser and W. Scott, *British Towns* (Oliver and Boyd, Edinburgh, 1961).

[14] E.H. Ruspini, A new approach to clustering, *Inform. and Control* **15** (1969) 22–32.

[15] S.Z. Selim, Using nonconvex programming techniques in cluster analyis, *Joint Meeting of the Operations Research Society of America and Institute of Management Science*, Houston (1981).

[16] S.Z. Selim, A global algorithm for the hard clustering problem, Submitted for publication, 1991.

[17] S.Z. Selim, A simulated annealing algorithm for the clustering problem, *Pattern Recognition* **24**(10) (1991).

[18] S.Z. Selim and M.A. Ismail, On the local optimality of the fuzzy Isodata clustering algorithm, *IEEE Trans. Pattern Anal. Machine Intelligence* **8**(2) (1986) 284–288.

[19] H. Späth, *Cluster Analysis Algorithms* (Ellis Horwood, Chichester, 1980).

[20] G.W. Stewart, *Introduction to Matrix Computations* (Academic Press, New York, 1973).

[21] W. Tucker, Counterexamples to the convergence theorem for fuzzy Isodata clustering algorithms, in: J. Bezdek, Ed., *The Analysis of Fuzzy Information.* (CRC Press, Boca Raton, IL, 1987) Vol. III, Ch. 7, 109–121.