sor in the Department of Computer Science, Wayne State University, Detroit, MI. In 1974, he joined the Department of Computer Science, Michigan State University, where he is currently a Professor. He served as the Program Director of the Intelligent Systems Program at the National Science Foundation from September 1980 to August 1981. His research interests are in the areas of pattern recognition and image processing.

Dr. Jain is a member of the Association for Computing Machinery, the Pattern Recognition Society, and Sigma Xi. He is also an Advisory Editor of *Pattern Recognition Letters.*

K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality

SHOKRI Z. SELIM AND M. A. ISMAIL, MEMBER, IEEE

Abstract—**The *K*-means algorithm is a commonly used technique in cluster analysis. In this paper, several questions about the algorithm are addressed. The clustering problem is first cast as a nonconvex mathematical program. Then, a rigorous proof of the finite convergence of the *K*-means-type algorithm is given for any metric. It is shown that under certain conditions the algorithm may fail to converge to a local minimum, and that it converges under differentiability conditions to a Kuhn–Tucker point. Finally, a method for obtaining a local-minimum solution is given.**

Index Terms—**Basic ISODATA, cluster analysis, *K*-means algorithm, *K*-means convergence, numerical taxonomy.**

I. Introduction

K-MEANS-type algorithms for exploratory data clustering and analysis are very popular and well known [1]-[8]. The main idea behind these techniques is the minimization of a certain criterion function usually taken up as a function of the deviations between all patterns from their respective cluster centers. Usually, the minimization of such a criterion function is sought utilizing an iterative scheme which starts with an arbitrary chosen initial cluster configuration of the data, then alters the cluster membership in an iterative manner to obtain a better configuration. The sum of squared Euclidean distances criterion has been adopted in most of the studies related to these algorithms, due to its computational simplicity, since the cluster at each iteration can be calculated in a straightforward manner. A *K*-means algorithm alternates between two major steps until a stopping criterion is satisfied. These steps are mainly the distribution of patterns among clusters utilizing a specific classifier (usually the minimum Euclidean distance classifier: MEDC), and the updating of cluster centers [9]-[11].

Incorporation of some heuristic procedures into the abovementioned iterative scheme results in the well-known ISODATA algorithm of Ball and Hall [12], which may be considered as another sophisticated form of the original *K*-means. The most important heuristic procedures in ISODATA are those allowing cluster lumping and cluster splitting.

Several extensive studies dealing with comparative analysis of different clustering methods have been conducted recently, utilizing both simulated data sets, e.g., [2], [13], and practical data, e.g., [6], [14]-[16]. These studies recommend the *K*-means algorithm as one of the best clustering methods available. Several evaluation criteria were adopted in such studies and a variety of techniques were compared simultaneously.

Moreover, fuzzy versions of the *K*-means algorithm have been reported in a series of papers by Ruspini [17], [18] Dunn [19], and Bezdek [19]-[22], where each pattern is allowed to have membership functions to all clusters rather than having a distinct membership to exactly one cluster.

The usefulness of the *K*-means-type algorithms is not questionable, and the extensive experimentation with these algorithms using practical data suggests and establishes the applicability and practicality of such techniques. Although it is found that such algorithms converge when applied to different data sets from a wide range of applications, no rigorous theorem for the convergence of the *K*-means-type algorithms exists to date, and the question of convergence of such methods remain open [1], [3], and [22].

In this paper, a rigorous proof of convergence of the *K*-means-type algorithm is given in a generalized form. Moreover, local optimality of solutions obtained has been investigated, where it is shown that under certain conditions, the *K*-means algo-

Manuscript received July 12, 1982; revised January 24, 1983. This work was supported by the University of Petroleum and Minerals, Dhahran, Saudi Arabia.

S. Z. Selim is with the Department of Systems Engineering, University of Petroleum and Minerals, Dhahran, Saudi Arabia.

M. A. Ismail is with the Department of Computer Science, University of Windsor, Windsor, Canada.

0162-8828/84/0100-0081$01.00

rithm may not yield local minimum solutions. In such cases, means of obtaining local minima are presented.

In Section II, the notation and mathematical formulation of the problem are established, where a general metric (or norm) is assumed to achieve generality. Different properties of the mathematical formulation are investigated in Section III. In Section IV, partial optimal solutions (POS's) are defined, and an algorithm for obtaining them is given. The convergence of the K-means algorithm is proven to be finite in Section V, and the conditions for local optimality are addressed in Section VI. In Section VII, the effect of the metric used on the local optimality of the POS's is explored, followed by Section VIII which contains a method of obtaining a local minimum solution in case the K-means algorithms fail to converge to one. Finally, the conclusions are reported in Section IX.

II. Mathematical Formulation of the Problem

Let $x_1, x_2, \cdots, x_m \in R^n$ be a finite number of patterns. To partition patterns into k partitions, $2 \leqslant k \leqslant m$, the following mathematical program is considered:

$$P\text{: minimize } f(W, Z) = \sum_{i=1}^{k} \sum_{j=1}^{m} w_{ij} D(x_j, z_i)$$

$$\text{subject to} \quad \sum_{i=1}^{k} w_{ij} = 1, \quad j = 1, 2, \cdots, m,$$

$$w_{ij} = 0 \text{ or } 1, \quad i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, m \qquad (1)$$

where

$W = [w_{ij}]$ is a $k \times m$ real matrix

$Z = [z_1, z_2, \cdots, z_k] \in R^{nk}$

$z_i \in R^n$ is the center of cluster i, $\quad i = 1, 2, \cdots, k$

$D(x_j, z_i)$ is some dissimilarity measure between x_j and z_i, $i = 1, 2, \cdots, k, j = 1, \cdots, m$.

Problem P is a nonconvex program where a local minimum point need not be global minimum. Cooper [23] proposed the above formulation where $D(x_j, z_i)$ was taken as the Euclidean distance.

A K-means-type algorithm could be interpreted in terms of the notation introduced above as follows.

1) Start with a set of initial cluster centers $z_i^{(1)}, i = 1, 2, \cdots, k$. Set $l = 1$.

2) Assign each pattern $x_j, j = 1, 2, \cdots, m$, to its nearest center which is equivalent to fixing the values of w_{ij}. Set $l = l + 1$.

3) Recompute the centers $z_i^{(l)}, i = 1, 2, \cdots, k$, by minimizing $F(\cdot, Z)$. If $z_i^{(l)} = z_i^{(l-1)}, i = 1, 2, \cdots, k$, stop; otherwise go to step 2).

In the following section, we will investigate some of the properties of Problem P above.

III. Properties of Problem P

To investigate the properties of Problem P we start by the following definition.

Definition 1: The reduced function $F(W)$ of Problem P is defined by

$$F(W) = \min \{f(W, Z): Z \in R^{nk}\},$$

and W is any $k \times m$ real matrix.

We note here that for some values of W, F is possibly unbounded.

Lemma 1: The reduced objective function F is concave.

Proof: Consider the two points W^1 and W^2 and let γ be any scalar such that $0 \leqslant \gamma \leqslant 1$; then

$$\begin{aligned} F(\gamma W^1 &+ (1 - \gamma) W^2) \\ &= \min \{f(\gamma W^1 + (1 - \gamma) W^2, Z): Z \in R^{nk}\} \\ &= \min \{\gamma f(W^1, Z) + (1 - \gamma) f(W^2, Z): Z R^{nk}\} \\ &\quad [\text{by linearity of } f(W, \cdot)] \\ &\geqslant \gamma \min \{f(W^1, Z): Z \in R^{nk}\} + (1 - \gamma) \\ &\quad \min \{f(W^2, Z): Z \in R^{nk}\} \\ &= \gamma F(W^1) + (1 - \gamma) F(W^2). \end{aligned}$$

Hence F is concave. Next we show an important property of the constraints set (1).

Theorem 2: Consider the set S given by

$$S = \left\{ \sum_{i=1}^{k} w_{ij} = 1, j = 1, 2, \cdots, m, w_{ij} \geqslant 0, \right.$$
$$\left. i = 1, 2, \cdots, k, j = 1, 2, \cdots, m \right\}.$$

The extreme points of S satisfy constraints (1).

Proof: Each extreme point of S is associated with a basis. Any basis of the constraints in S is an identity matrix. Hence, each basic variable will have value 1 and nonbasic variables will be zeros. This completes the proof.

Next we define an important problem which is equivalent to Problem P.

Definition 2: The Reduced Problem RP of problem P is given by

$$\text{minimize } F(W), \text{ subject to } W \in S.$$

Since the function F is concave then there exist an extreme point solution of Problem RP which in turn satisfies constraints (1). The following lemma is immediate.

Lemma 3: Problems RP and P are equivalent.

In the rest of the paper any results pertaining to one of the problems will transmit to the other. At this point we mention an interesting fact. If in the formulation of Problem P, we let $w_{ij} \in [0, 1]$, then the solution of Problem P will satisfy $w_{ij} = 0$ or 1. Moreover, if the function $f(W, Z)$ is concave in W for fixed Z then still the optimal value of $w_{ij} = 0$ or 1. On the other hand, if $f(W, Z)$ is convex in W, then the optimal values of w_{ij} may be fractional giving rise to fuzzy clustering [17], [20].

As noted earlier Problem P has local minimum points. In the next section we characterize points which are called partial optimal solutions and show that they are Kuhn-Tucker points of Problem P. (See [25] for definition and significance of Kuhn-Tucker points.)

IV. Partial Optimal Solutions and Kuhn-Tucker Points

Definition 3: A point (W^*, Z^*) is partial optimal solution of problem P if it satisfies the following [24]:

$$f(W^*, Z^*) \leqslant f(W, Z^*) \qquad \text{for all } W \in S,$$

and

$$f(W^*, Z^*) \leqslant f(W^*, Z) \qquad \text{for all } Z \in R^{nk}.$$

To obtain a partial optimal solution we define the following two problems:

Problem P_1: Given $\hat{Z} \in R^{nk}$, minimize $f(W, \hat{Z})$ subject to $W \in S$.

Problem P_2: Given $\hat{W} \in S$, minimize $f(\hat{W}, Z)$ subject to $Z \in R^{nk}$. The solution of Problem P_1 is trivial since for a specific j,

$$w_{rj} = 1 \qquad \text{if } D(X_j, z_r) \leqslant D(X_j, z_l),\ l = 1, 2, \cdots, k$$

and $w_{ij} = 0$ for $i \neq r$. However the solution of Problem P_2 may not be as straightforward as P_1. Note that (W^*, Z^*) is a partial optimal solution of P if W^* solves P_1 when $\hat{Z} = Z^*$ and Z^* solves P_2 when $\hat{W} = W^*$. In the next theorem we show that partial optimal solutions are Kuhn-Tucker points of Problem P.

Theorem 4: Suppose $F(W^*, Z)$ is differentiable at $Z = Z^*$, then (W^*, Z^*) is a Kuhn-Tucker point of Problem P if and only if it is a partial optimal solution P.

Proof: Let W_j be the jth column of W, i.e.,

$$W_j = (w_{ij}, w_{2j}, \cdots, w_{kj})', \qquad j = 1, 2, \cdots, m$$

$$\nabla_{w_j} f(W, Z) = \{\cdots \partial f(W, Z)/\partial w_{ij} \cdots\}, \qquad j = 1, 2, \cdots, m$$

and

$$\nabla_{z_i} f(W, Z) = \{\partial f(W, Z)/\partial z_{i1}, \cdots, \partial f(W, Z)/\partial z_{in}\}.$$

Note that $S = \{(W_1, W_2, \cdots, W_m) \mid E\, W_j = 1, W_j \geqslant 0, j = 1, \cdots, m\}$ where $E \in R^n$ is a vector of ones.

Hence, Kuhn-Tucker conditions are given by

$$\text{i)}\ \nabla_{w_j} f(W, Z) + \Pi^j E \geqslant 0 \qquad j = 1, \cdots, m$$

$$\text{ii)}\ (\nabla_{w_j} f(W, Z) + \Pi^j E)\, W_j = 0 \qquad j = 1, \cdots, m$$

$$\text{iii)}\ E\, W_j = 1,\ W_j \geqslant 0 \qquad j = 1, \cdots, m$$

$$\text{iv)}\ \nabla_{z_i} f(W, Z) = 0 \qquad i = 1, \cdots, k.$$

Now assume (W^*, Z^*) is a partial optimal solution satisfying the differentiability condition of the theorem. Since W^* solves P_1 for $\hat{Z} = Z^*$ then W^* satisfies Kuhn-Tucker conditions of P_1 which are given by i), ii), and iii) when $Z = Z^*$. Also if Z^* solves P_2 for $\hat{W} = W^*$ then Z^* satisfies its Kuhn-Tucker condition given by iv) where $W = W^*$. Hence (W^*, Z^*) satisfy Kuhn-Tucker conditions of problem P. Now assume (W^*, Z^*) is a Kuhn-Tucker point of P. Suppose that (W^*, Z^*) is not a partial optimal solution of P, i.e., W^* does not solve Problem P_1 for $Z = Z^*$. If this is true, then W^* should not solve Kuhn-Tucker conditions of Problem P_1 for $Z = Z^*$ which are given by i), ii), and iii). This is a contradiction and hence (W^*, Z^*) is a partial optimal solution. This completes the proof.

Although a partial optimum solution is a Kuhn-Tucker point of Problem P it may not even be a local minimum. The following example illustrates this possibility. Let the patterns be $X_1 = (0, 0)$, $X_2 = (4, 0)$, $X_3 = (4, 2)$, and $X_4 = (8, 2)$ and $k = 2$. If the K-means algorithms is used with initial cluster centers $Z_1 = (2, 0)$ and $Z_2 = (6, 2)$ then it will stop after one iteration with clusters (X_1, X_2) and (X_3, X_4) and $W_{11}^* = W_{12}^* = W_{23}^* = W_{24}^* = 1$, $W_{ij} = 0$ otherwise, and $f(W^*, Z^*) = 8$ where Euclidean distances are used. However, this point is not a local minimum since if we define $\hat{W} = \{W_{12} = 1 - \epsilon,\ W_{22} = \epsilon$ and all other W_{ij}'s are the same as before, $\epsilon > 0$ and small$\}$, then $\min f(\hat{W}, Z) = 6$ and the optimal centers are $(0, 0)$ and $(4, 2)$.

In the next section an algorithm for obtaining partial optimal solutions is given and is shown to converge finitely.

V. Convergence of K-Means-Type Algorithms

The following algorithm generates partial optimal solutions. It is essentially a restatement of the K-means algorithm.

Algorithm 1:

i) Choose an initial point $Z^\circ \in R^{nk}$, solve P_1 with $Z = Z^\circ$. Let W° be an optimal basic solution of P_1. Set $r = 0$.

ii) Solve P_2 with $\hat{W} = W^r$. Let the solution be Z^{r+1}. If $f(\hat{W}, Z^{r+1}) = f(\hat{W}, Z^r)$ stop; set $(W^*, Z^*) = (W^*, Z^{r+1})$; otherwise go to step iii).

iii) Solve P_1 with $\hat{Z} = Z^{r+1}$, let the basic solution be W^{r+1}, if $f(W^{r+1}, \hat{Z}) = f(W^r, \hat{Z})$ stop; set $(W^*, Z^*) = (W^{r+1}, \hat{Z})$; otherwise let $r = r + 1$ and go to step ii).

Theorem 5: Algorithm 1 converges to a partial optimal solution of Problem P in a finite number of iterations.

Proof: First we show that an extreme point of S, is visited at most once by the algorithm before it stops. Assume that this is not true, i.e., $W^{r_1} = W^{r_2}$ for some r_1, r_2 where $r_1 \neq r_2$. When applying step ii) we get Z^{r_1+1} and Z^{r_2+1} as optimal solutions for $W = W^{r_1}$ and $W = W^{r_2}$, respectively, i.e.,

$$\begin{aligned} f(W^{r_1}, Z^{r_1+1}) &= f(W^{r_2}, Z^{r_1+1}), \qquad \text{since } W^{r_1} = W^{r_2} \\ &= f(W^{r_2}, W^{r_2+1}), \qquad \text{from step ii).} \end{aligned} \tag{2}$$

But the sequence $f(\cdot, \cdot)$ generated by the algorithm is strictly decreasing; hence (2) is not true and $W^{r_1} \neq W^{r_2}$. Since there are a finite number of extreme points of S, then the algorithm will reach a partial optimal solution in a finite number of iterations. This completes the proof.

As concluded in the previous section the K-means algorithm may yield a point which is not a local minimum of Problem P. In the following section, we discuss local optimality of the partial optimal point.

VI. Local Optimality of the Partial Optimal Solution

In this section, we make the valid assumption that there is a closed and bounded set $V \in R^{nk}$ where $Z^* \in V$. A well-known result is that a center z_i^* lies in the convex hull of the patterns forming cluster i. Hence the smallest such V is the union of the convex hulls of all clusters. Another characterization of V is the convex hull of all patterns.

In the following lemma an important result is given.

Lemma 6: Let $f(W, Z)$ be defined for all W and for all $Z \in V$.

Assume that:

i) V is compact and

ii) f and the partial derivatives $\partial f / \partial w_{ij}$ are continuous.

Let $F(W) = \min \{f(W, Z): Z \in V\}$ and

$$A(W^*) = \{Z: Z \text{ minimizes } f(W^*, Z), Z \in V\}. \tag{3}$$

Let the (one-sided) directional derivatives of F at W^* in the direction d be given as

$$DF(W^*; d) = \lim_{\alpha \to 0^+} F(W^* + \alpha d) - F(W^*)/\alpha.$$

Then $DF(W^*; d)$ exists for any d at any point W^* and is given by

$$DF(W^*; d) = \min \{\nabla_w'(W^*, Z) \cdot d: Z \in A(W^*)\}$$

where $\nabla_w(W^*, Z)$ is the vector of partial derivatives

$$\partial f(W, Z)/\partial w_{ij} \qquad \text{evaluated at } W = W^*.$$

Proof: See [25, Theorem 1, p. 420].

A well-known optimality condition is given in the following lemma.

Lemma 7: Consider Problem *RP*: $\min \{F(W): W \in S\}$, and S is convex. Let $W^* \in S$. Then W^* is a local minimum of *RP*, if and only if the directional derivatives $DF(W^*; d) \geqslant 0$ for each feasible direction d at W^*.

The next theorem characterizes the local optimality of the partial optimal solution obtained by the K-means algorithm (Algorithm 1).

Theorem 8: Let (W^*, Z^*) be a given point such that W^* is an extreme point of the polyhedral set S and $Z^* \in A(W^*)$. Then W^* is a local minimum of Problem *RP* if and only if

$$F(W^*) = f(W^*, Z^*)$$
$$\leqslant \min \{f(W, Z): W \in S \text{ for all } Z \in A(W^*)\}. \tag{4}$$

Proof: Consider Problem *RP* and assume (4) holds. Consider some $\hat{Z} \in A(W^*)$, by assumption

$$f(W^*, Z^*) = f(W^*, \hat{Z}) \leqslant \min \{f(W, \hat{Z}): W \in S\}.$$

This implies that for any feasible direction d at W^*

$$\nabla_w' f(W^*, \hat{Z}) \cdot d \geqslant 0.$$

Since this is true for any $Z \in A(W^*)$,

$$\min \{\nabla_w' f(W^*, Z) \cdot d: Z \in A(W^*)\} \geqslant 0.$$

But from Lemma 6 the left-hand side is $DF(W^*; d)$. Hence by Lemma 7, W^* is a local minimum of Problem *RP*. Now assume W^* is a local minimum of *RP*. Then for any feasible direction d we must have by Lemma 7, $DF(W^*; d) \geqslant 0$. By Lemma 6 $DF(W^*: d) = \min \{\nabla_w' f(W^*, Z) \cdot d: Z \in A(W^*)\}$. Hence,

$$\nabla_w' f(W^*, Z) \cdot d \geqslant 0 \qquad \text{for each } Z \in A(W^*). \tag{5}$$

Now consider a fixed $\hat{Z} \in A(W^*)$. $f(W, Z^*)$ is linear in W, and (5) is true for all feasible directions. This implies $f(W^*, \hat{Z}) \leqslant \min \{f(W, \hat{Z}): W \in S\}$ but $\hat{Z} \in A(W^*)$ is arbitrary. Hence, (4) holds. This completes the proof.

If the set $A(W^*)$ is singleton, characterization of the local minimum point becomes very simple as shown in the following theorem.

Theorem 9: Let (W^*, Z^*) be a partial optimal solution of Problem P and let $A(W^*)$ given by (3) be singleton; then W^* is a local minimum of Problem *RP*.

Proof: Since W^* is a partial optimal solution, then by Definition 3, $f(W^*, Z^*) \leqslant \min \{f(W, Z^*): W \in S\}$. But this is precisely the condition of local optimality of W^* if $A(W^*)$ is singleton (Theorem 8). Hence the proof is complete.

In the next section we investigate the effect of the dissimilarity function $D(x_j, z_i)$ on local optimality.

VII. Metrics and Local Optimality

In this section we consider two metrics:

i) Quadratic metric: $D(x_j, z_i) \triangleq (x_j - z_i)'(x_j - z_i)$, and

ii) Minkowsky metric: $D(x_j, z_i) \triangleq \left(\sum_{l=1}^{n} |x_{jl} - z_{il}|^P\right)^{1/P}.$

In the second class, if $P = 1$ the rectilinear norm is obtained while if $P = 2$ we obtain the Euclidean norm. Finally, for $P \to \infty$ we obtain Chebyshev norm. The purpose of the following analysis is to identify the conditions pertaining to each metric which cause $A(W^*)$ to be singleton, hence by Theorem 8 a partial optimal solution will be a local minimum of Problem *RP*.

To simplify the analysis we let $f(W, Z) = \Sigma_{i=1}^{k} f_i(W_i, z_i)$ where W_i is the ith row of the matrix W and

$$f_i(W_i, z_i) = \sum_{j=1}^{m} w_{ij} D(x_j, z_i).$$

Let $A(W^*)$ be as defined by (3) and

$$A_i(W_i^*) = \{z_i: z_i \text{ minimizes } f_i(W_i^*, z_i),$$
$$W_i^* \text{ is the } i\text{th row of } W^*, z_i \in V_i\}$$

and $V_i \in R^n$ is a compact set which contains the optimal center z_i (e.g., V_i could be the convex hull of the patterns of cluster i). Obviously, the set $A(W^*)$ is singleton if and only if each of the sets $A_i(W_i^*)$, $i = 1, 2, \cdots, k$ is singleton. In the next theorem we give the conditions that $A_i(W_i^*)$ is nonsingleton.

Theorem 10: Let $D(x_j, z_i)$ be defined by a Minkowsky metric. Let W_i^* and $A_i(W_i^*)$ be as defined above. Then the set $A_i(W_i^*)$ is nonsingleton if and only if:

i) the points x_j, $j = 1, 2, \cdots, m$ are collinear and

ii) $\Sigma_{j=1}^{m} w_{ij}^*$ is even.

Proof: To show sufficiency, assume that conditions i) and ii) hold. From condition i); $x_j = a + b\, t_j$ where $a, b \in R^n$ are given and some $t_j \in R$. The solution z_i^* is in the convex hull of the patterns x_j forming the cluster; hence we are interested in the points z_i, where $z_i = a + b\,\theta_i$. Let $b' = (b_1, b_2, \cdots, b_n)'$; hence,

$$D(x_j, z_i) = \left(\sum_{l=1}^{n} |x_{jl} - z_{il}|^P\right)^{1/P}$$

$$= \left(\sum_{l=1}^{n} b_l^P |\theta_i - t_j|^P\right)^{1/P}$$

$$= |\theta_i - t_j| \left(\sum_{l=1}^{n} b_l^P\right)^{1/P}.$$

Now

$$f_i(W_i^*, z_i) = \sum_{j=1}^{m} w_{ij}^* D(x_j, z_i)$$

$$= \left(\sum_{j=1}^{m} W_{ij}^* |\theta_i - t_j|\right) \left(\sum_{l=1}^{n} b_l^P\right)^{1/P}.$$

Hence, the problem of finding the optimal vector z_i has been reduced to that of finding the optimal value of the scalar θ_i. The latter problem is the well-known problem of locating a point on a straight line [23]. We will assume without loss of generality that $t_1 \leqslant t_2 \leqslant \cdots \leqslant t_m$. Then it can be shown [26, p. 181] that the optimal solution $\theta_i^* = t_r$ for some index r, where

$$\sum_{j=1}^{r-1} w_{ij}^* < \sum_{j=r}^{m} w_{ij}^* \text{ and } \sum_{j=1}^{r} w_{ij}^* \geqslant \sum_{j=r+1}^{m} w_{ij}^*. \tag{6}$$

x_r is said to be a median location.

Since $w_{ij} = 0$ or 1 then $\Sigma_{j=1}^{m} w_{ij}$ is integer. By condition ii) $\Sigma_{i=1}^{m} w_{ij}^* = 2M$, where M is an integer scalar. Let r be such that $\Sigma_{j=1}^{r} w_{ij}^* = M$; then the optimal solution is $\theta_i^* = t_r$ or alternatively $\theta_i^* = t_{r+1}$ as one can easily verify that inequalities (6) are satisfied by both points t_r and t_{r+1}. Moreover, by the convexity of the function $f_i(W_i^*, z_i)$, each point on the line segment $[t_r, t_{r+1}]$ is an optimal solution. Thus $A_i(W_i^*)$ is nonsingleton.

To show necessity, assume $A_i(W_i^*)$ is nonsingleton and suppose conditions i) and ii) do not hold. The function $f_i(W_i^*, z_i)$ is a sum of strictly convex functions hence f_i is strictly convex and it has a unique minimum which is a contradiction. This completes the proof.

If quadratic metrics are used, then partial optimal solutions are always local minimum points as shown in Theorem 11.

Theorem 11: Consider Problem P where $D(x_j, z_i) = (x_j - z_i)'(x_j - z_i)$; then a partial optimal solution of P is a local minimum point.

Proof: Let (W^*, Z^*) be a partial optimum solution of P; then $Z^* = (z_1^*, z_2^*, \cdots)$ and $z_i^* = \Sigma_{j=1}^{n} w_{ij}^* x_j / \Sigma_{j=1}^{n} w_{ij}^*$ which is a unique value hence $A_i(W_i^*)$ is singleton as well as $A(W^*)$. By Theorem 9 the proof is complete.

In the following section we discuss how to obtain a local minimum point of Problem P in case the K-means algorithm fails to converge to one.

VIII. Attaining Local Minimum Points

Let (W^*, Z^*) be a partial solution of Problem P: recall the definition of S and define the set

$$T(W^*) = \{W \in R^{nk}: W \text{ is an extreme point of } S \text{ and } W \text{ is adjacent to } W^*\}.$$

Note that if two extreme points of a polyhedral set are adjacent then they correspond to two bases of the constraints which differ in exactly two variables. Now we introduce the following result.

Theorem 12: Let $\overline{W} \in S$ be such that $F(\overline{W}) \leqslant F(W)$ for all $W \in T(\overline{W})$. Then $\overline{W}$ is a local minimum points of Problem RP.

Proof: Let d be any feasible direction of S at $\overline{W}$ and d^q, $q \in Q$, be the set of extreme directions of S incident to the extreme point $\overline{W}$. Then $d = \Sigma_{q \in Q}\, \mu_q d^q$ for some $\mu_q \geqslant 0$, $q \in Q$.

Now

$$DF(\overline{W}; d) = \min \{\nabla'_w f(\overline{W}, Z) \cdot d: Z \in A(\overline{W})\}$$

$$= \min \left\{ \sum_{q \in Q} \mu_q \nabla'_w f(\overline{W}, Z) \cdot d^q: Z \in A(\overline{W}) \right\}$$

$$= \sum_{q \in Q} \mu_q \min \{\nabla'_w f(\overline{W}, Z) \cdot d^q: Z \in A(\overline{W})\}$$

$$= \sum_{q \in Q} \mu_q DF(\overline{W}; d^q). \tag{7}$$

Since $F(\overline{W}) \leqslant F(W)$ for all $W \in T(\overline{W})$, then $DF(\overline{W}; d^q) \geqslant 0$ for all $q \in Q$. By substituting in (7) we get $DF(\overline{W}; d) \geqslant 0$ for any feasible direction and since S is convex then by Lemma 7, the result follows. This completes the proof.

Now suppose the K-means algorithm stops at the point (W^*, Z^*) where W^* is not a local minimum of Problem RP. Then one could be reached by examining the points $W^q \in T(W^*)$, $q \in Q$. If some point $W^s \in T(W^*)$, $s \in Q$ satisfies $F(W^s) < F(W^*)$ (W^s always exists) then Algorithm 1 (K-means algorithm) could be used starting with step ii).

In the following we investigate the process of examining the extreme points of the set $T(W^*)$. To obtain an element of $T(W^*)$ a variable which is nonbasic is chosen to become basic and it replaces a basic variable at the extreme point W^*.

The total number of variables w_{ij} is mk. At any extreme point of S exactly m variables are basic (each equal unity). Hence, there are $m(k - 1)$ nonbasic variables and consequently the cardinality of the set T is $m(k - 1)$.

Suppose at the point W^*, we have $w_{eg}^* = 0$ and it is decided to make w_{eg} basic. The variable which will become nonbasic is w_{lg} where $w_{lg}^* = 1$. This results in a new basis W^r for some r, where $w_{eg}^r = 1$, $w_{lg}^r = 0$ and $w_{ij}^r = w_{ij}^*$ for all other values of i and j.

Now

$$F(W^r) = \min \sum_{i=1}^{k} f_i(W_i^r, z_i): z_i \in R^n, i = 1, 2, \cdots, k$$

where W_i^r is the ith row of the matrix W^r

$$\begin{aligned} F(W^r) &= \sum_{i=1}^{k} \min_{z_i} f_i(W_i^r, z_i) \\ &= \sum_{\substack{i=1 \\ l \neq i \neq e}}^{k} \min_{z_i} f_i(W_i^r, z_i) + \min_{z_l} f_l(W_l^r, z_l) \\ &\quad + \min_{z_e} f_e(W_e^r, z_e). \end{aligned} \tag{8}$$

Note that

$$W_i^r = W_i^* \qquad \text{for } i = 1, 2, \cdots, k, e \neq i \neq l.$$

Hence

$$\begin{aligned} F(W^r) = &\sum_{\substack{i=1 \\ e \neq i \neq l}}^{k} \min_{z_i} f_i(W_i^*, Z_i) + \min_{z_l} f(W_l^r, z_l) \\ &+ \min_{z_e} f_e(W_e^r, z_e). \end{aligned}$$

The first term is known from previous computation. So to compute $F(W^r)$ only two centers have to be recomputed, namely z_l and z_e. Essentially we are considering a specific pattern g, release it from cluster l and allocate it to cluster e, recompute the new centers and compare $F(W^r)$ with $F(W^*)$. Let us call clusters l and e the releasing and receiving clusters respectively. Consider the releasing cluster l, it can easily be verified that

$$\min_{z_l} f_l(W_l^*, z_l) \geqslant \min_{z_l} f_l(W_l^r, z_l) + D(x_g, z_l^*). \tag{9}$$

For the receiving cluster e the following is true:

$$\min_{z_e} f_e(W_e^*, z_e) + D(x_g, z_e^*) \geqslant \min_{z_e} f_e(W_e^r, z_e). \tag{10}$$

From (8), (9), and (10) we get

$$\begin{aligned} F(W^r) - F(W^*) &= \min_{z_l} f_l(W_l^r, z_l) + \min_{z_e} f_e(W_e^r, z_e) \\ &\quad - \min_{z_l} f_l(W_l^*, z_l) - \min f_e(W_e^*, z_e) \\ &\leqslant D(x_g, z_e^*) - D(x_g, z_l^*). \end{aligned} \tag{11}$$

The right-hand side of (11) is always nonnegative (see solution to Problem P_1 Section IV), and as it gets smaller the higher the chances that $F(W^r) < F(W^*)$.

Based on (11) and the above comment heuristics could be developed to obtain an improving extreme point $W^r \in T(W^*)$ quickly. For example, compute the values of the right-hand side of (11) for $e = 1, 2, \cdots, k$, $g = 1, 2, \cdots, m$, and $w_{eg}^* = 0$. Arrange these values in an ascending order. Then investigate the extreme points corresponding to the arranged values in a sequential manner until W^r is reached.

IX. Conclusion

In this paper the clustering problem is formulated as a mathematical program. Properties of the latter were investigated and an equivalent concave program was constructed. Partial optimal solutions were defined and an algorithm for obtaining them is given which is basically a restatement of K-means algorithm. We have shown that POS's are Kuhn-Tucker points if differentiability is satisfied. If Minkowsky metric is used then under certain conditions the POS's may not even be a local minimum of the problem. For the latter case we describe how to obtain a local minimum of the problem. Currently, research work is underway to find the global solution [27].

References

[1] R. Dubes and A. K. Jain, "Cluster methodologies in exploratory data analysis," *Advances in Comput.*, vol. 19, pp. 113-228, 1980.

[2] —, "Clustering techniques: The user's dilemma," *Pattern Recognition*, vol. 8, pp. 247-260, 1976.

[3] J. Tou and R. Gonzales, *Pattern Recognition Principles.* Reading, MA: Addison-Wesley, 1974.

[4] E. Diday and J. C. Simon, "Cluster Analysis," in *Digital Pattern Recognition*, K. S. Fu, Ed. Berlin: Springer-Verlag, 1976.

[5] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic, 1972.

[7] J. A. Hartigan, *Clustering Algorithms.* New York: Wiley, 1975.

[8] M. R. Anderberg, *Cluster Analysis for Applications.* New York: Academic, 1973.

[9] E. Forgy, "Cluster analysis of multivariable data: Efficiency versus interpretability of classifications," *Biometrics*, vol. 21, p. 768 (abstr.), 1965.

[10] R. C. Jancey, "Multidimensional group analysis," *Austral. J. Botany*, vol. 14, no. 1, pp. 127-130, 1966.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariable observations," in *Proc. 5th Symp. Math. Statist. and Probability*, Berkeley, CA, vol. 1, AD 669871. Berkeley, CA: Univ. California Press, 1967, pp. 281-297.

[12] G. H. Ball and D. J. Hall, "ISODATA–A novel method of data analysis and pattern classification," Stanford Res. Inst., Menlo Park, CA, AD 699616, 1965.

[13] G. W. Milligen, "An experimentation of the effect of six types of error perturbation on fifteen clustering algorithms," Working Paper Series, College of Administrative Sci., Ohio State University, Columbus, 1978.

[14] J. E. Mezzich, "Evaluating clustering methods for psychiatric-diagnosis," *Biol. Psychiatry*, vol. 13, pp. 265-281, 1978.

[15] A. J. Boyce, "Mapping diversity: A comparative study of some numerical methods," in *Numerical Taxonomy*, A. J. Cole, Ed. New York: Academic, 1969, pp. 1-31.

[16] E. Filsinger and W. J. Sauer, "Empirical typology of adjustment to aging," *J. Gerontol.*, vol. 33, pp. 437-445, 1978.

[17] E. R. Ruspini, "A new approach to clustering," *Inform. Contr.*, vol. 15, pp. 22-32, 1969.

[18] —, "New experimental results in fuzzy clustering," *Inform. Sci.*, vol. 6, pp. 273-284, 1973.

[19] J. C. Dunn, "Some recent investigations of a fuzzy algorithm and its application to pattern classification problems," Center for Appl. Math., Cornell Univ., Ithaca, NY, 1974.

[20] J. C. Bezdek, "Fuzzy mathematics in pattern classification," Ph.D. dissertation, Center for Appl. Math., Cornell Univ., Ithaca, NY, 1973.

[21] —, "A physical interpretation of fuzzy ISODATA," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, pp. 387-390, 1976.

[22] —, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, pp. 1-8, 1980.

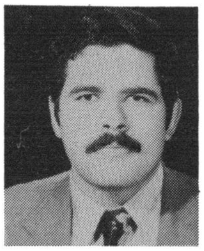
[23] L. Cooper, "M-dimensional location models: Application to cluster analysis," *J. Regional Sci.*, vol. 13, pp. 41-54, 1973.

[24] R. Wendel and A. Hurter, "Minimization of a non-separable objective function subject to disjoint constraints," *Oper. Res.*, vol. 24, pp. 643-657, 1976.

[25] L. S. Lasdon, *Optimization Theory for Large Systems.* New York: Macmillan, 1970.

[26] R. Francis and J. White, *Facility Layout and Location: An Analytical Approach.* Englewood Cliffs, NJ: Prentice-Hall, 1974.

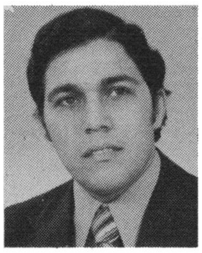
[27] S. Z. Selim, "Using nonconvex programming techniques in cluster analysis," presented at the ORSA/TIMS Joint Meeting, Houston, TX, Oct. 1981.



Shokri Z. Selim was born in Alexandria, Egypt, in 1948. He received the B.S. degree in mecanical engineering and the M.Sc. degree in industrial engineering from the University of Cairo, Cairo, Egypt, and the Ph.D. degree in operations research from Georgia Institute of Technology, Atlanta, in 1970, 1973, and 1979, respectively.

In 1979 he joined the Department of Systems Engineering, University of Petroleum and Minerals, Dhahran, Saudi Arabia, as an Assistant Professor. His research interests are in the areas of cluster analysis, simulation of large systems, location-allocation problems, and parameters estimation.

Dr. Selim is an associate member of ORSA and a senior member of IIE.



M. A. Ismail (S'75–M'79) was born in Alexandria, Egypt, on June 4, 1948. He received the B.Sc. (Honors) and M.Sc. degrees in electrical engineering and computer science from Alexandria University, Alexandria, Egypt, in 1970 and 1974, respectively, and the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, Ont., Canada, in 1979.

He taught at different Universities including the University of Waterloo, Waterloo, Ont., Canada; University of Petroleum and Minerals, Dhahran, Saudi Arabia; Alexandria University, Alexandria, Egypt, and University of Windsor, Windsor, Ont., Canada, where he is now an Associate Professor of Computer Science. His current research interests include pattern recognition, data structures and analysis, fuzzy clustering algorithms, machine intellignece, mathematical modeling, and computer applications in medicine, especially computer-aided diagnosis and prognosis.

Dr. Ismail is a member of the Pattern Recognition Society, Association for Computing Machinery, Society for Computer Simulation and several IEEE societies.

Correspondence

EMERGE—A Data-Driven Medical Decision Making Aid

DONNA L. HUDSON AND THELMA ESTRIN

Abstract—**EMERGE is an expert system designed as a medical decision making aid. It is machine-independent, and is implemented in standard Pascal. It has modest memory requirements, and can operate on a microcomputer. EMERGE is rule-based, and its initial application is the analysis of chest pain in the emergency room. The knowledge base is maintained separately from the consultation program. Thus the application area can be changed without any modification to the software. This paper describes the control structures and rule searching procedures used in EMERGE.**

Index Terms—**Artificial intelligence, decision support, emergency room procedures, expert system, medical decision making, microcomputer applications, rule-based system.**

Introduction

EMERGE, a rule-based expert system, has the following characteristics.

1) *Domain Independence:* The application area can be changed by replacing the rule base, with no modification required to the consultation program.

2) *Machine-Independence:* EMERGE is written in standard Pascal. It is conservative of memory, permitting its use on a microcomputer.

3) *Input Flexibility:* The user can enter information in free phrases, and can enter as little or as much information as desired. EMERGE obtains missing information by asking questions.

4) *Data-Driven Search:* The data-driven approach removes the necessity of the user answering questions which are not relevant to the case at hand.

The purpose of this paper is to describe the control structures and rule searching procedures used in EMERGE.

Analysis of existing expert systems reveals several areas where further investigation is necessary. These include development of machine-independent systems, methods for acquisition of adequate knowledge bases, development of appropriate control structures in the program, and adaptation of expert system technology to microcomputers [2], [3], [14], [16].

A major design consideration for EMERGE was to make it machine-independent and capable of running on microcomputers [8]. These requirements necessitate modifications in the user interface and control structures which are not necessary in rule-based programs implemented on large computer systems. Most rule-based systems are written in Lisp. Although Lisp is now available on some microcomputers, its use of memory is not efficient enough to allow implementation of a large rule-based system. In addition, the versions of Lisp available are not completely standardized. The language commonly available for microcomputers, Basic, was also considered. An example of a simple rule-based demonstration program written in Basic for an Apple microcomputer is given by Duda *et al.* [4]. The authors point out the inadequacy of Basic for this type of application, and also note the difficulties of implementing an expert system which deals with realistic problems on a microcomputer. Pascal was chosen as the language for EMERGE, since it is available on most microcomputers, it is standardized, and it allows recursive procedures which facilitate the efficient implementation of rule searches.

The initial implementation area for EMERGE is the evaluation of chest pain in the emergency room [5], [9]. Accurate decision making in this area is crucial. Seriously ill patients

Manuscript received November 20, 1981; revised December 27, 1982 and May 17, 1983.

D. L. Hudson is with the School of Medicine, University of California, Fresno-Central San Joaquin Valley Medical Education Program, Fresno, CA 93703.

T. Estrin is with the Department of Computer Science, University of California, Los Angeles, CA 90024.

0162-8828/84/0100-0087$01.00 © 1984 IEEE