FUZZY CLUSTERING WITH A FUZZY COVARIANCE MATRIX

Donald E. Gustafson and William C. Kessel
Scientific Systems, Inc.
186 Alewife Brook Parkway
Cambridge, Massachusetts 02138

Abstract

A class of fuzzy ISODATA clustering algorithms has been developed previously which includes fuzzy means. This class of algorithms is generalized to include fuzzy covariances. The resulting algorithm closely resembles maximum likelihood estimation of mixture densities. It is argued that use of fuzzy covariances is a natural approach to fuzzy clustering. Experimental results are presented which indicate that more accurate clustering may be obtained by using fuzzy covariances.

1. Introduction

The notion of fuzzy sets, first put forth by Zadeh [1], is an attempt to modify the basic conception of a space--that is, the set on which the given problem is defined. By introducing the concept of a fuzzy--i.e., an unsharply defined set, a different perspective is provided for certain problems in systems analysis, including pattern recognition.

One of the significant difficulties in development of a systematic approach to pattern recognition is that the phenomena of interest are modeled by equations which contain functions and operators which may appear simple and natural, but which yield some solutions which could be regarded as pathological. The difficulty stems from our desire to differentiate between classes in a manner which is simple and easy to visualize. In doing so, we restrict the solutions in an unknown way. The use of fuzzy sets is an attempt to ameliorate this problem.

Pattern classification problems have provided impetus for the development of fuzzy set theory. Recently, fuzzy sets have provided a theoretical basis for cluster analysis with the introduction of fuzzy clustering. The use of fuzzy sets in clustering was first proposed in [2] and several classification schemes were developed [3]. The first fuzzy clustering algorithm was developed in 1969 by Ruspini [4], and used by several workers [5]. Following this, Dunn [6] developed the first fuzzy extension of the least-squares approach to clustering and this was generalized by Bezdek [7] to an infinite family of algorithms.

Several problems in medical diagnosis have been attacked using fuzzy clustering algorithms. Adey [8] achieved promising results in interpreting EEG patterns in cerebral systems. Bezdek [9] has studied its use in differentiating hiatal hernia and gallstones. It appears that medical diagnosis may be an especially fruitful area of application for fuzzy clustering, since biological systems are extremely complex and the boundaries between "distinct" medical diagnostic classes are not sharply defined. This has been suggested for cardiovascular investigations [10].

In a "hard" clustering algorithm, each pattern vector must be assigned to a single cluster. This "all or none" membership restriction is not a realistic one, since many pattern vectors may have the characteristics of several classes. It is more natural to assign to each pattern vector a set of memberships, one for each class. The implication of this is that the class boundaries are not "hard" but rather are "fuzzy". Another problem is that the set of all partitions resulting from a "hard" clustering algorithm is extremely large, making an exhaustive search extremely complicated and expensive. Fuzzy clustering will generally lead to more computational tractability [11]. Another advantage of fuzzy clustering is that troublesome or outlying members of the data set are more easily recognized than with hard clustering, since the degree of membership is continuous rather than "all-or-none." Bezdek and Dunn [12] have noted the relationship of fuzzy clustering to estimating mixture distributions, but retained the Euclidean metric. Here, a generalization to a metric which appears more natural is made, through the use of a <u>fuzzy covariance matrix</u>.

2. Problem Formulation

The definition of a fuzzy partition used here agrees with that of Ruspini [4], Dunn [6] and Bezdek [13] and is a natural extension of the conventional partitioning definition. An ordinary, or "hard" partition is a k-tuple of Boolean functions $w(\cdot) = \{w_1, w_2, \ldots, w_k\}$ on the feature space $\Gamma \subset R^n$ which satisfy

$$w_j(x) = 0 \text{ or } 1, \ \forall\, x \in \Gamma, \ 1 \leq j \leq k \quad (1)$$

$$\sum_{j=1}^{k} w_j(x) = 1 \quad \forall\, x \in \Gamma \quad (2)$$

If Γ_j represents the j-th class, with $\Gamma_i \cap \Gamma_j = \Phi \ \forall\, i \neq j$ and $\cup_{j=1}^{k} \Gamma_j = \Gamma$, then $w_m(x) = 1$ means that $x \in \Gamma_m$ and (2) insures that x is a member of

CH1392-0/78/0000-0761$00.75 © 1978 IEEE

precisely one class. It is possible to pass from this definition to a corresponding fuzzy partition by retaining (2) but replacing (1) with the relaxed condition $0 \le w_j \le 1$. Thus, a fuzzy partition is a k-tuple of membership functions $w(\cdot) = \{w_1(x), w_2(x), \ldots, w_k(x)\}$ which satisfy

$$0 \le w_j(x) \le 1, \ \forall x \varepsilon \Gamma, \ 1 \le j \le k \tag{3}$$

$$\sum_{j=1}^{k} w_j(x) = 1, \quad \forall x \ \varepsilon \ \Gamma \tag{4}$$

Equation (3) suggests a probabilistic interpretation for the membership functions, as discussed by Ruspini [4]. However, this may or may not be a correct interpretation.

In devising a conventional clustering algorithm, one typically looks for a scalar performance index which attains its minimum for a partition which maximally separates the naturally-occurring clusters. There should exist a feasible algorithm for minimizing the performance index. The weighted within-class squared error is a useful performance measure.

Denote the distance from a point x to the j-th class by $d_j(x) = d(x, \theta_j)$; $d_j(x) > 0$, where the j-th class is parametrized by θ_j. For an indexed set of samples $x_1, x_2, x_3, \ldots, x_N$ we denote the distance measure and membership function by $d_j(x_i) = d_{ij}$, $w_j(x_i) = w_{ij}$. We are interested in minimizing the following cost:

$$J(w,\theta) = \sum_{i=1}^{N} \sum_{j=1}^{k} w_{ij}^{\ \alpha} d_{ij} \ ; \ \alpha \ge 1 \tag{5}$$

where $\theta = \{\theta_j\}$, $w = \{w_{ij}\}$, k is the number of classes, and α is a smoothing parameter which controls the "fuzziness" of the clusters. For $\alpha = 1$, the clusters are separated by hard partitions and w_{ij} = 0 or 1. As α increases, the partitions become more fuzzy.

3. Determination of Fuzzy Clusters

3.1 Determination of Optimal Membership Functions

Now consider the problem of minimizing J with respect to (fuzzy)w, subject to $\alpha>1$ and the constraints (3) and (4). We defer for later the determination of the optimal parameters by minimizing J over θ. Constraint (3) may be eliminated by setting $w_{ij} = S_{ij}^2$ with S_{ij} real. We adjoin the constraints (3) and (4) to J with a set of Lagrange multipliers $\{\lambda_i\}$ to give

$$\bar{J}(S,\theta,\lambda) = \sum_{i=1}^{N} \sum_{j=1}^{k} S_{ij}^{\ 2\alpha} d_{ij} + \sum_{i=1}^{N} \lambda_i \left(\sum_{j=1}^{k} S_{ij}^{\ 2} - 1 \right) \tag{6}$$

The first-order necessary conditions for optimality are found by setting the gradients of $\bar{J}$ with respect to S to zero. Now,

$$\frac{\partial \bar{J}}{\partial S_{ij}} = 2\alpha \ S_{ij}^{\ 2\alpha-1} d_{ij} + 2S_{ij} \ \lambda_i \tag{7}$$

By setting $\frac{\partial \bar{J}}{\partial S}$ to zero we obtain the following first-order necessary conditions:

$$S^*_{ij}(\alpha S^{*\ 2(\alpha-1)}_{ij} d_{ij} + \lambda^*_i) = 0; \ \forall \ i,j \tag{8}$$

$$\sum_{j=1}^{k} S^{*\ 2}_{ij} = 1 \qquad ; \ \forall \ i \tag{9}$$

where the asterik denotes association with optimality.

Equations (8) - (9) comprise a set of NK + N equations which can be solved for the Nk + N unknowns $W^* = \{w_{ij}^*\}$, and $\lambda^* = \{\lambda^*_i\}$. We proceed by first assuming that $S^*_{ij} \ne 0 \ \forall i, j$. This is consistent with the assumption that $\alpha > 1$. With this assumption we have

$$w^*_{ij} = (-\lambda^*_i / \alpha d_{ij})^{1/(\alpha-1)} \tag{10}$$

By summing over j and using (4)

$$(-\lambda^*_i)^{\frac{1}{\alpha-1}} = \frac{1}{\sum_{j=1}^{k} \left(\frac{1}{\alpha d_{ij}}\right)^{1/(\alpha-1)}} \tag{11}$$

and (10) becomes

$$w^*_{ij} = \frac{1}{\sum_{\ell=1}^{k} (d_{ij}/d_{i\ell})^{1/(\alpha-1)}} \tag{12}$$

Then, from (5), <u>for any θ</u>, the associated extremum of $J(w,\theta)$ is

$$J^*(\theta) = \min_{w} J(w,\theta)$$

$$= \sum_{i=1}^{N} \left[\sum_{j=1}^{k} (d_{ij})^{1/(1-\alpha)} \right]^{1-\alpha} \tag{13}$$

<u>Limiting Case When $\alpha \to 1$</u>

If $\alpha \to 1$,

$$J \to \sum_{i=1}^{N} \sum_{j=1}^{k} w_{ij} \ d_{ij} \tag{14}$$

and the argument given by Dunn [6] will establish that $\forall$ i,k

$$w^*_{ik} \to \begin{cases} 1; \ d_{ik} = \min_{j} (d_{jk}) \\ 0; \ \text{otherwise} \end{cases} \tag{15}$$

provided $\min_{j}(d_{jk})$ is unique $\forall$k. Otherwise, W* is a hard k-partition which is unique up to arrangements caused by tie-breaking rules.

3.2 Determination of Optimal Parameters

We now turn to the problem of finding the optimal parameter set $\theta^* = \{\theta^*_1, \theta^*_2, \ldots, \theta^*_k\}$. From (5) we have

$$\frac{\partial}{\partial \theta_j} \bar{J}(w,\theta,\lambda) = \sum_{i=1}^{N} w_{ij}^{\ \alpha} \ \frac{\partial}{\partial \theta_j} d_{ij} \tag{16}$$

The first-order necessary conditions for a local minimum of J are (8), (9) and

$$\sum_{i=1}^{N} w^{*\ \alpha}_{ij} \frac{\partial}{\partial \theta_j} d_{ij} \Big|_* = 0 \ \forall j \tag{17}$$

To proceed we need to specify the parametrization of d_{ij}.

<u>Fuzzy ISODATA</u>. Let $d_{ij} = (x_i - \theta_j)^T A (x_i - \theta_j); \ A>0$ (18)

Then (17) gives

$$\sum_{i=1}^{N} w^{*\ \alpha}_{ij} (x_i - \theta^*_j) = 0 \ \forall j \tag{19}$$

This is equivalent to

$$\theta_j^* = \frac{\sum_{i=1}^{N} w_{ij}^{*\,\alpha} x_i}{\sum_{i=1}^{N} w_{ij}^{*\,\alpha}} \triangleq m_{fj};\ k=1,\ldots,k \tag{20}$$

We will call m_{fj} the fuzzy mean of class j in recognition of its limiting property under hard partitioning. This case comprises fuzzy ISODATA [14].

Hard ISODATA. As $\alpha \to 1$ and the partitioning becomes hard:

$$w_{ij}^{*\,\alpha} \quad \begin{cases} 1 \ ; \ j=m \\ 0 \ ; \ j \neq m \end{cases} \tag{21}$$

where

$$d_{im} = \min_j d_{ij} \tag{22}$$

That is, under the one-nearest-neighbor rule, $w_{ij}^{*\,\alpha}\big|_{\alpha=1} = 1$ for all pattern vectors x_i assigned to class j and is zero otherwise. Thus, for hard partitioning

$$\sum_{i=1}^{N} w_{ij}^* = N_j \tag{23}$$

where N_j is the number of pattern vectors assigned to Γ_j and

$$\theta_j^*\big|_{\alpha\to 1} \to \frac{1}{N_j} \sum_{x_i \in \Gamma_j} x_i = \hat{m}_j \tag{24}$$

where $\hat{m}_j$ is the sample mean of Γ_j. This is the hard k-means algorithm: it constitutes the basic idea underlying hard ISODATA [15].

3.3 Generalization to Include Fuzzy Covariance

Now consider replacing (18) by an inner product induced norm metric of the form

$$d_{ij}(\theta_j) = (x_i - v_j)^T M_j (x_i - v_j), \quad 1 \le j \le k \tag{25}$$

with M_j symmetric and positive-definite. If $\theta_j = v_j$, equation (20) for θ_j^* still holds [14]. If, however, we take $\theta_j = \{v_j, M_j\}$, a class of algorithms more general than fuzzy ISODATA will ensue. Note that J is now linear in M_j, giving a singular problem. The cost J may be made as small as desired by simply making M_j less positive definite. To get a feasible solution, we must constrain M_j in some manner. Ideally we would like the metric to handle different scalings along each direction in feature space. That is, we would like to allow variations in the shape of each class induced by the metric but not let the metric grow without bound. A way of accomplishing this by using only one parameter is to constrain the determinant $|M_j|$ of the matrix M_j. This induces a volume constraint.

Consider the set of constraints

$$|M_j| = \rho_j \ , \ \rho_j > 0 \tag{26}$$

with ρ_j fixed for each j. The augmented cost is now

$$J(w,\theta,\lambda,\beta) = \sum_{i=1}^{N} \sum_{j=1}^{k} w_{ij}^{\alpha} d_{ij}(\theta_j) + \sum_{i=1}^{N} \lambda_i \left(\sum_{j=1}^{k} w_{ij} - 1 \right) + \sum_{j=1}^{k} \beta_j \left(|M_j| - \rho_j \right) \tag{27}$$

where $\{\beta_j\}$ is a set of Lagrange multipliers.

The partial derivatives with respect to θ_j now change. From (27), the necessary conditions are

$$\frac{\partial J}{\partial v_j}\bigg|_* = -2 \sum_{i=1}^{N} w_{ij}^{\alpha} M_j (x_i - v_j^*) = 0 \ ; j=1,2,\ldots,k \tag{28}$$

which is identical to (19) and

$$\frac{\partial J}{\partial M_j}\bigg|_* = 0 = \sum_{i=1}^{N} w_{ij}^{\alpha} (x_i - v_j)(x_i - v_j)^T + \beta_j |M_j^*| M_j^{*-1} \tag{29}$$

To get (29), we have used the identities

$$\frac{\partial}{\partial A}(x^T A x) = xx^T \quad , \quad \frac{\partial}{\partial A}|A| = |A| A^{-1}$$

which hold for a non-singular matrix A and any compatible vector x. Eq. (28) gives (20) again:

$$v_j^* = \frac{\sum_{i=1}^{N} w_{ij}^{\alpha} x_i}{\sum_{i=1}^{N} w_{ij}^{\alpha}} \tag{30}$$

For the optimal membership functions $(w_{ij} = w_{ij}^*)$, v_j^* is the fuzzy mean of Γ_j. Eq. (29) gives, for $v_j = v_j^*$,

$$M_j^{*-1} = \frac{1}{\beta_j |M_j^*|} \sum_{i=1}^{N} w_{ij}^{\alpha} (x_i - v_j^*)(x_i - v_j^*)^T \tag{31}$$

Now define the fuzzy covariance matrix for Γ_j by

$$P_{fj} = \frac{\sum_{i=1}^{N} w_{ij}^{\alpha} (x_i - m_{fj})(x_i - m_{fj})^T}{\sum_{i=1}^{N} w_{ij}^{\alpha}} \ ; \ \alpha > 1 \tag{32}$$

Then, using (32) and (26) in (31) gives

$$M_j^{*-1} = \left(\frac{1}{\rho_j |P_{fj}|} \right)^{1/n} P_{fj} \tag{33}$$

where n is the feature space dimension. In the sequel, a hard covariance matrix refers to P_{fj} of (32) evaluated at $\alpha=1$. In view of (21), a hard covariance matrix is simply the sample class covariance matrix under the cluster assignment rule (22).

The previous discussion suggests the following iterative algorithm for finding stationary points of $J(w,\theta)$. Given data $\{x_i\}$ and an initial guess $\theta_j^{(o)} = \{m_{fj}^{(o)}, P_{fj}^{(o)}\}$, we proceed as follows:

for k=1,2,... :

(i) compute $\{d_i(\theta_j^{(k)})\}$ using (25).

(ii) compute $\{w_{ij}^{(k)}\}$ using (12). If $d_{ik}=0$ for some k, set $w_{ik}=1$, $w_{i\ell}=0 \ \forall \ell \neq k$.

(iii) compute new estimates $\theta_j^{(k+1)}$ using (30), (32) and (33). Recycle to (i) until a specified convergence criterion is satisfied.

4. Relation to Maximum Likelihood Estimation

There is an intimate relationship between fuzzy ISODATA algorithms and maximum likelihood algorithms designed to estimate mixture density parameters under the Gaussian assumption. Maximum likelihood estimation of parameters has been studied for a long time (see, e.g., Rao, 1952[16]), and the theory is quite well understood. The problem in applications is developing numerical techniques which can efficiently solve, or approximately solve, the problem. The development here follows the work of Wolfe [17].

Let $p(x|\Gamma_j)$ be the probability density for the random vector $x \epsilon R^n$, conditioned on x being a member of the j-th class ($x \epsilon \Gamma_j$), and let P_j be the a priori probability associated with Γ_j. We assume that Γ_j is parametrized by a set of parameters $\theta_j \epsilon R^s$ and that $p(x|\Gamma_j)$ is a twice differentiable function of θ_j. Since x can be associated with more than one class, it has a mixture density function which is, for k classes,

$$p(x) = \sum_{j=1}^{k} P_j p(x,\theta_j) \ , \quad \sum_{j=1}^{k} P_j = 1 \qquad (34)$$

where $p(x,\theta_j) = p(x|\Gamma_j)$. The "probability of membership" of x in class j can be found by using Bayes' Rule:

$$p(\Gamma_j|x) = \frac{P_j p(x,\theta_j)}{p(x)} \qquad (35)$$

Now suppose a sample of N random vectors is drawn from the mixture and denote these by $x_1,x_2,x_3,\ldots x_N$. Then, assuming independent sampling, the log probability is $\log p(x_1,x_2,\ldots,x_N) = \sum_{i=1}^{N} \log p(x_i)$. The maximum likelihood estimate of the parameters $\theta=\theta_1,\theta_2,\ldots,\theta_k$ is found by solving $\max_\theta[\log p(x_1,x_2,\ldots,x_N)]$ subject to the constraint in (34). The first order necessary conditions are

$$P_j^* = \frac{1}{N}\sum_{i=1}^{N} p^*(\Gamma_j|x_i) \qquad (36)$$

$$\sum_{i=1}^{N} p^*(\Gamma_j|x_i)\frac{\partial}{\partial\theta_j}\log p^*(x_i,\theta_j^*) = 0 \qquad (37)$$

Now consider the special case where x is conditionally Gaussian distrubuted. Then

$$\log p(x,\theta_j) = -\frac{n}{2}\log 2\pi + \tfrac{1}{2}\log|E_j^{-1}| - \tfrac{1}{2}(x-m_j)^T E_j^{-1}(x-m_j) \qquad (38)$$

where $\theta_j=\{m_j,E_j\}$ and E_j is assumed nonsingular. Taking the indicated partial derivatives in (37), we obtain the following three equations which describe the necessary conditions to be satisfied for the maximum likelihood estimates

$$m_j^* = \frac{1}{NP_j^*}\sum_{i=1}^{N} p(x_i,\theta_j^*)x_i \ , \quad P_j^* = \frac{1}{N}\sum_{i=1}^{N} p(x_i,\theta_j^*) \qquad (39)$$

$$E_j^* = \frac{1}{NP_j^*}\sum_{i=1}^{N} p(x_i,\theta_j^*)(x_i-m_j^*)(x_i-m_j^*)^T \qquad (40)$$

The first order necessary conditions for fuzzy clustering and maximum likelihood estimation possess similarities which can be studied by imbedding both solutions in a larger class of solutions. Consider the following set of algebraic relations:

$$Q_j = \frac{1}{N}\sum_{i=1}^{N} q_{ij} \ , \quad \eta_j = \frac{1}{NQ_j}\sum_{i=1}^{N} q_{ij}x_i \ ; \quad 0 \le q_{ij} \le 1$$

$$M_j = \frac{\gamma_j}{NQ_j}\sum_{i=1}^{N} q_{ij}r_{ij}r_{ij}^T \ , \quad r_{ij} = x_i - \eta_j \text{ with } x_i \epsilon R^n,$$

$\eta_j \epsilon R^n$, N a positive integer, and γ_j a positive scalar. The parameter q_{ij} is the membership function of x_i relative to class j and Q_j is the average membership for class j. Thus, q_{ij} increases as x_i comes closer to class j and relatively large values of Q_j are associated with the largest or most dense classes. The parameter η_j can be regarded as the nucleus point of class j and M_j is a matrix which describes the shape and size of the class. The parameter r_{ij} is the vector from x_i to the class j nucleus. The parameters r_{ij}, M_j are combined into a measure d_{ij} which is used to evaluate the distance x_i to class j: $d_{ij} = r_{ij}^T M_j^{-1} r_{ij}$

The values of q_{ij} and the associated constraints for fuzzy clustering and maximum likelihood estimation are summarized in Table 1. The parameter D_i is a normalization constant for x_i and C_j is a normalization constant for Γ_j. Note that q_{ij} decreases monotonically with increasing d_{ij} for both cases. It is also interesting to note that membership functions are normalized differently. With fuzzy clustering, normalization is done over the classes to get D_i, whereas normalization under maximum likelihood estimation is done over the whole space R^n to obtain C_j. Thus, q_{ij} is given a slightly different interpretation in the two methods. The constraints are quite different: a class volume constraint is used under fuzzy clustering whereas a total probability constraint is used under maximum likelihood estimation.

Even with these differences, there is a striking similarity between the two methods. Note in particular that the fuzzy covariance matrix appears naturally in the problem and appears to be more appropriate than a hard covariance matrix.

We now consider how to build a classifier using the q_{ij}'s from either maximum likelihood or fuzzy ISODATA. The decision rule by which x_i is assigned to a class is as follows:

Assign x_i to class m if $\quad q_{im} \ge q_{ij}$; $j=1,2,\ldots,k$

In case of ties, assign x_i to the least-numbered class.

5. Fuzzy Clustering Experiments

The fuzzy clustering algorithm has been implemented and tested using two stylized classes which had some degree of overlap. The two classes are depicted in Figure 1 and consist of two long and narrow regions at right angles to one another in a cross pattern. The two cluster centroids coincide exactly so that the discrimination must be based on cluster shape information. In order to test the algorithms, a total of ten points in each class were chosen randomly, using a uniform distribution over each class. These points are depicted in Figure 1, with points labeled x selected from Class 1 and points labeled o selected from Class 2. All tests were run assuming two classes apriori. Updating of the covariance matrices was done using either: (a) full updating (use (32) directly in (25)), (b) no updating (use initial guess at all steps, (c) $|M_j|$ = constant (i.e., invoke (26)). The iterations were stopped when the change in each membership function was less than 0.001 in magnitude.

A test was run using hard ISODATA (α=1,A=I) seeded with the sample means. The resulting assignments are shown in Figure 2 and are poor since class shape is not accounted for. The algorithm converged after only two passes. The next test used fuzzy IDODATA, in which the means were fuzzy but A=I. The resulting clusters are shown in Figure 3 and are considerably different from the desired result.

Cluster 1 is very large and Cluster 2 is very small, encompassing only three peripheral points of Class 1. Convergence was obtained in 4 passes.

A test was next run using fuzzy clustering with α=2 and using fuzzy covariance matrices, with initial guesses $M_{f1}^{(o)} = M_{f2}^{(o)} = I$. The clusters were seeded at the sample class means. The class assignments are shown in Figure 5 and are seen to be correct for all points, although the results for #5 and #11 would appear fortuitous. The difficulty in classifying these two points is apparent from the values of their membership functions. Thirteen passes were required to meet the convergence criterion.

The next run was similar to the previous run except that the cluster seeds were set at S_1=(0.001, 0), S_2=(0,0) which were used in the fuzzy ISODATA run, in order to make the discrimination more difficult. The discrimination was, in fact, more difficult. However, after 20 passes, the algorithm did converge to the configuration of Figure 6. As before, all of the assignments were correct. However, the way in which the clusters were formed was quite interesting. The histories of the membership functions for several critical points are given in Table 2 and demonstrate the nature of the iterative process. Note that $w_{3,1}$, $w_{4,1}$, $w_{10,1}$, $w_{13,2}$, and $w_{19,2}$ increase monotonically and approach a value of unity. This is the desired behavior and is expected for points which lie much closer to one class than the other. Note that the response of $w_{13,2}$ is relatively slow, staying close to 0.5 until the 15th pass and then increasing monotonically. Thus, for the first 14 passes point #13 is about equally distant from both clusters. Point #11 is strongly associated with Cluster 1 on the 11th through 15th passes. However, once point #10 is correctly assigned to Cluster 1, $w_{11,2}$ increases monotonically to its final value. Note that points #5 and #11 are both strongly associated with Cluster 1 from the 11th through 15th pass, indicating that Cluster 2 does not start to form correctly until the 16th pass.

The effect of using a fuzzy covariance matrix was studied by running a case differing from the previous one only in the way the covariance matrix was calculated. A hard covariance matrix was used instead of a fuzzy one. The solution is shown in Figure 8 and was obtained after eight passes. Note that points #4 and #11 are incorrectly classified. The failure to correctly assign #11 is hardly surprising but the misassignment of #4 is judged to be a clustering error. This result suggests that the use of fuzzy covariances can enhance clustering performance. Further numerical testing is required to verify this behavior in general.

It is interesting to note that the configuration of Figure 8 is relatively insensitive to the distance measure used. A run was made in which the distance measure $1 - \exp(-d_{ij}/2)$ was used rather than d_{ij} and the same cluster assignments were obtained. It should also be noted that no problems of convergence were encountered in any runs.

REFERENCES

1. L.A. Zadeh, "Fuzzy Sets", Information and Control, Vol. 8, pp. 338-353, 1965.
2. R.E. Bellman, R.A. Kalaba and L.A. Zadeh, "Abstraction and Pattern Classification", J. Math. Anal. Appl., Vol. 13, pp. 1-7, 1966.
3. I. Gitman and M. Levine, "An Algorithm for Detecting Unimodal Fuzzy Sets and Its Application as a Clustering Technique", IEEE Trans. Computers, Vol. C-19, pp. 917-923, 1970.
4. E.H. Ruspini, "A New Approach to Clustering", Information and Control, Vol.15, pp.22-32, 1969.
5. L. Larsen, E. Ruspini, J. McDew, D. Walter and W. Adey, "A Test of Sleep Staging Systems in the Unrestrained Chimpanzee", Brain Research, Vol. 40, pp. 319-343, 1972.
6. J. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", J. Cybernetics, Vol. 3, pp. 32-57, 1974.
7. J. Bezdek, "Fuzzy Mathematics in Pattern Classification", Ph.D. Thesis, Cornell University, Ithaca, New York, 1973.
8. W. Adey, "Organization of Brain Tissue: Is the Brain a Noisy Processor?", Int. J. Neuroscience, Vol. 3, pp. 271-284, 1972.
9. J.C. Bezdek, "Feature Selection for Binary Data-Medical Diagnosis with Fuzzy Sets", National Computer Conference, 1976.
10. D. Kalmanson, H.F. Stegall, "Cardiovascular Investigations and Fuzzy Sets Theory", Amer. J. of Cardiology, Vol. 35, pp. 30-34, 1975.
11. J. Bezdek, "Cluster Validity with Fuzzy Sets", J. Cybernetics, Vol. 3, pp. 58-73, 1974.
12. J.C. Bezdek and J.C. Dunn, "Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Distributions", IEEE Trans. Comp. Vol C-24, pp. 835-838, 1975.
13. J.C. Bezdek, "Numerical Taxonomy with Fuzzy Sets", J. Math. Biology, Vol. 1, pp.57-71, 1974.
14. J.C. Bezdek and P.F. Castelaz, "Prototype Classification and Feature Selection with Fuzzy Sets", IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-7, No. 2, February, 1971, pp. 87-92.
15. G.H. Ball, "Classification Analysis", Stanford Res. Inst. Rept. AD-716-482, November, 1970.
16. C.R. Rao, Advanced Statistical Methods in Biometric Research, New York: Wiley and Sons, 1952.
17. J.H. Wolfe, "Pattern Clustering by Multivariate Mixture Analysis", Multivariable Behavioral Research, July, 1970, pp. 329-350.

Fuzzy Clustering	parameter or condition	Maximum Likelihood
w_{ij}^{α} $(\alpha \geq 1)$	$q_{ij}=q_j(x_i)$	P_{ij}
$w_{ij}=d_{ij}^{1/(1-\alpha)}/D_i$		$P_{ij}=C_j \exp[-d_{ij}/2]$
$\sum_{j=1}^{k} w_{ij}=1 \Rightarrow D_i \ \forall i$	normalization	$\int_{X\epsilon R^n} p(x)dx=1 \Rightarrow C_j \ \forall j$
$\lvert M_j \rvert=\rho_j \Rightarrow \gamma_j$	constraints	$\sum_{j=1}^{k} Q_j=1$

Table 1 Comparison of Fuzzy Clustering and Maximum Likelihood Solutions

Pass	$W_{3,1}$	$W_{4,1}$	$W_{5,1}$	$W_{10,1}$	$W_{11,2}$	$W_{13,2}$	$W_{14,2}$	$W_{19,2}$
1	.5001	.5004	.5015	.5001	.4993	.5000	.5000	.5000
7	.5029	.5001	.5153	.5002	.4907	.4996	.5119	.5022
10	.5218	.5495	.7047	.5085	.3676	.4934	.6599	.5341
15	.8690	.9574	.8664	.8757	.0104	.6992	.9353	.9373
17	.9905	.9521	.7921	.9899	.2424	.9680	.9808	.9905
18	.9972	.9509	.6850	.9968	.6268	.9911	.9794	.9965
20	.9988	.9606	.6836	.9985	.7403	.9949	.9715	.9975

Table 2 Membership Function Histories for Case Shown in Figure 5.

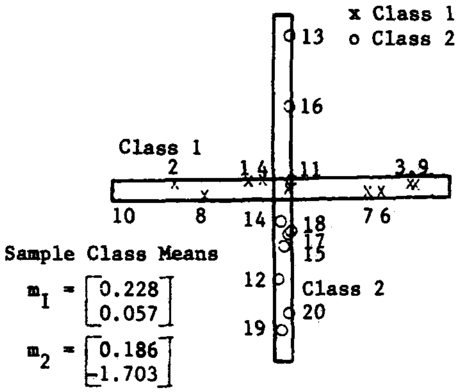

Figure 1: Two-Class Configuration

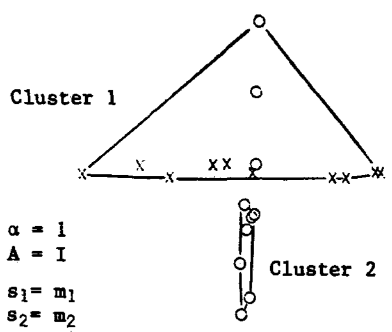

Figure 2: Cluster Assignments Using Hard ISODATA Seeded With Class Sample Means

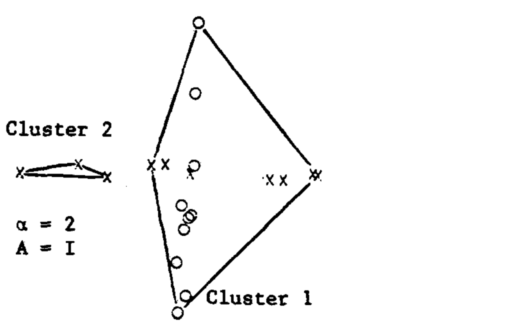

Figure 3: Cluster Assignments Using Fuzzy ISODATA With Seeds $S_1 = (0.001,0)$, $S_2 = (0,0)$

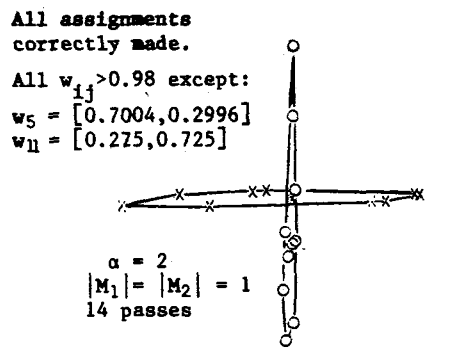

Figure 4: Cluster Assignments Using Fuzzy Covariance Seeded at Class Means

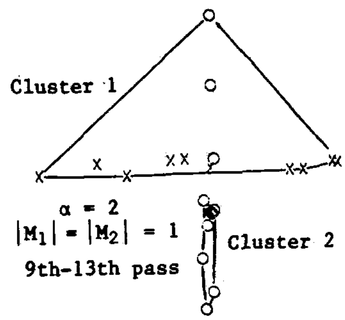

Figure 5(a): Cluster Assignments Using Fuzzy Covariance With Seeds $S_1 = (0.002,0)$, $S_2 = (0,0)$

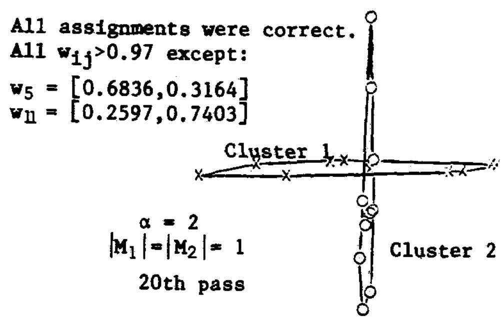

Figure 5(b): Cluster Assignments Using Fuzzy Covariances With Seeds $S_1 = (0.001,0)$, $S_2 = (0,0)$ After Convergence

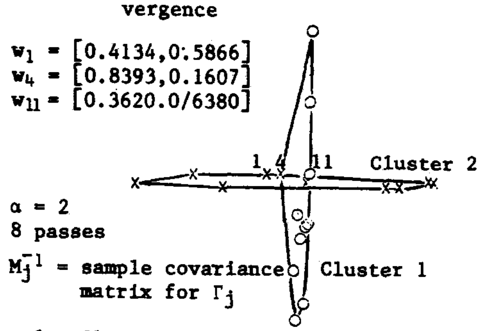

Figure 6: Cluster Assignments Using Fuzzy ISODATA Seeded at $S_1 = (0.001,0)$, $S_2 = (0,0)$ and Sample Covariance Matrices