Feature selection for binary data—Medical diagnosis with fuzzy sets*

by JAMES C. BEZDEK
Marquette University
Milwaukee, Wisconsin

ABSTRACT

The notion of fuzzy sets—sets with imprecise boundaries—is a natural cornerstone upon which to build algorithms based on approximate reasoning. Since their inception in 1965 by Zadeh,[1] fuzzy sets have led to first steps towards quantifying data analysis in many fields previously immune to mathematical examination. A fairly exhaustive introduction to the theory and applications of fuzzy sets[2] lists 238 papers dealing with a great variety of recent investigations.

INTRODUCTION

In this paper we discuss the applicability of the fuzzy ISODATA clustering algorithms for (1) dimensionality reduction of binary valued data sets, and (2) computerized medical diagnosis. The first question is often referred to as feature selection; overviews of many popular approaches are available in References 3 and 4. Loosely speaking, one wants to reduce the number of characteristics originally measured to some optimal subset which retains at least as much information about substructure in the data as the original ones. Computerized medical diagnosis is an extremely difficult and ambitious undertaking. Considering the risks involved, enormous improvements need to be made in existing methodologies before the medical community can be asked to rely on the diagnostic suggestions of a computer. However, it is our conviction that the attitude of pessimism displayed in Reference 5 towards this enterprise is largely attributable to the failure of conventional (that is, non-fuzzy) techniques, the results of which must either be accepted at face value or rejected out of hand: we believe that fuzzy sets can be used as a basis for computerized diagnostic *advice* that will provide valuable insight and direction for clinicians with a large data base of previous case histories.

* This research supported by National Science Foundation Grant DCR75-05014.

FUZZY CLUSTERING ALGORITHMS

Let R^d denote real, d-dimensional Euclidean space (*feature space*), and let $X=\{x_1,x_2,\ldots,x_n\}\subset R^d$. Each $x_k=(x_{k1},x_{k2},\ldots,x_{kd})\in R^d$ is a *feature vector* (subject, patient); each x_{kj} in R is the j^{th} *feature* (characteristic, attribute, symptom) of feature vector x_k; and if every $x_{kj}\in\{0,1\}$, we call X a binary valued data set. In this instance, we say x_k has attribute j when $x_{kj}=1$, and is lacking it if $x_{kj}=0$. Cluster analysis with respect to X is the problem of finding an integer c, $2\leq c<n$, and c subsets (clusters) of X which partition it into subgroups of points revealing intrinsic substructure in the data. Algorithms to partition X abound; the partitions they find depend on the classification criterion used by the algorithm which defines similarity between pairs of vectors in X.

There are *hard* (i.e., conventional) and fuzzy methods, and each of these main classes can be roughly subdivided into graph-theoretic and objective function techniques. Readers interested in hard algorithms for clustering will find an introduction to the literature in Reference 3; a brief review of fuzzy clustering follows. Clustering with fuzzy sets was first proposed in Reference 6. References 7-10 discuss some of the earliest fuzzy pattern classification schemes. In 1969 Ruspini delineated the first fuzzy clustering method based on objective functions, and foreshadowed the usefulness of information measures (entropy) in the fuzzy sets context. His technique was enlarged and illustrated in References 12-15. Dunn[16] defined the first fuzzy extension of the classical within group sum of squared errors (WGSS) objective functional, and in Reference 17 this approach was generalized to yield the infinite family of algorithms discussed below. Methods of clustering based on fuzzy graphs are still in their infancy; References 18-23 are seminal works in this direction.

CLUSTERING WITH FUZZY ISODATA

A hard c-partition P of X is a collection of c nonempty subsets of X, say $P=\{Y_1,Y_2,\ldots,Y_c\}$, whose

union is X and whose pairwise intersections are disjoint. To characterize P by a matrix, let u_i: X→{0, 1} be the characteristic function of Y_i:

$$u_i(x_k) = u_{ik} = \left\{ \begin{matrix} 1 \text{ in case } x_k \in Y_i \\ 0 \text{ otherwise} \end{matrix} \right\}; 1 \leq i \leq c; 1 \leq k \leq n. \quad (1)$$

Denote by V_{cn} the vector space of all real c×n matrices, let $U \in V_{cn}$, and let u_{ik} be the ik^{th} entry of U. The set of matrices

$$M_c = \left\{ U \epsilon V_{cn} : u_{ik} \epsilon \{0,1\} \; \forall \; i,k; \sum_{i=1}^{c} u_{ik} = 1 \; \forall \; k; \sum_{k=1}^{n} u_{ik} > 0 \; \forall \; i \right\} \quad (2)$$

is called *hard c-partition space* for X because each partition P of X corresponds uniquely to the matrix in M_c whose rows are the values for the characteristic functions of the subsets in P as shown in Equation (1).

Solutions for all hard clustering algorithms lie in M_c, and this is a fundamental drawback for two reasons: First, each member of the data must be assigned unequivocal membership in one and only one of the c partitioning subsets; however, the substructure in real data rarely—if ever—is so distinct that every member in X is most realistically described as a full member of a single subclass. A fuzzy model can overcome this objection by allowing every individual partial membership in all c subsets, as, for example, one would desire for hybrids when classifying them in parallel with their progenitors. Secondly, M_c is a finite but extremely large set, a complication which often manifests itself in analytical as well as computational intractabilities.

Fuzzy sets provide a natural way to surmount the objections above. We call any *function* u_i that maps X into the closed interval [0,1] a *fuzzy subset* or fuzzy cluster in X. The number $u_i(x_k) = u_{ik}$ is the *grade of membership* of subject x_k in fuzzy set u_i, and fuzzy c-partitions of X are defined by imbedding M_c in

$$M_{fc} = \left\{ U \epsilon V_{cn} : u_{ik} \epsilon [0,1] \; \forall \; i,k; \sum_{i=1}^{c} u_{ik} = 1 \; \forall k; \sum_{k=1}^{n} u_{ik} > 0 \; \forall \; i \right\}. \quad (3)$$

M_{fc} is called *fuzzy c-partition space* associated with X. The requirement that each column in U sum to one stipulates that every vector in X be assigned a *total* membership equal to unity in the partitioning subsets. If M_c is enlarged to include matrices which may have some zero rows, say M_{co}, then M_{fc} is the convex hull of M_{co}.[17] Compactness, convexity, and continuity endow M_{fc} with a pleasant mathematical structure; for example, it has been shown[24] numerically that because M_{fc} is continuous, algorithms defined on it have paths of feasible solutions around undesirable local trap states of algorithms confined to M_c.

Given M_{fc}, how can fuzzy c-partitions of X be found? One way to identify optimal fuzzy clusterings in X is via the family of generalized WGSS error objective functionals defined in Reference 17. On the Cartesian product of M_{fc} with R^{cd}, we define for $m \epsilon [1, \infty)$

$$J_m(U,v) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m ||x_k - v_i||^2. \quad (4)$$

In (4) $U \epsilon M_{fc}$, $v = (v_1, v_2, \ldots, v_c) \epsilon R^{cd}$, $v_i = (v_{i1}, v_{i2}, \ldots, v_{id}) \epsilon R^d$ for $1 \leq i \leq c$, and $||\cdot||$ is any norm on feature space. J_m is an etxension of the classical minimum variance objective functional J_1 because $J_m = J_1 \; \forall \; m$ whenever $U \epsilon M_c$ is hard. The c vectors $\{v_i\}$ comprising v are presumed to have features prototypical of vectors in X having a high affinity for membership in the respective fuzzy clusters $\{u_i\}$, and so are called *cluster centers* of their respective fuzzy clusters. These vectors will play an important role in the sequel. The measure of similarity in (4) is the norm $||\cdot||$; in this model it compares members of the data to each other indirectly via distances between them and the cluster centers.

Optimal fuzzy c-partitionings of X are defined as part of solution pairs $(\hat{U}, \hat{v})$ of the optimization problem

$$\text{minimize}\{J_m(U,v)\} \quad \text{over } M_{fc} \otimes R^{cd}. \quad (5)$$

Partitions arising as part of solutions for (5) are related to a well defined type of hard, compact, well separated (CWS) clusters for X in Reference 16. There are structured data sets whose clusters do not enjoy this property, but there is a wide class of patterns for which this criterion is very basic, and we adopt it here as implicit in our clustering goals.

Necessary conditions for solutions of (5) were derived[17] for the class of functionals in (4) whose norms were differentiable (e.g., inner product induced norms). It was shown there that for m>1 and $x_k \neq \hat{v}_i \; \forall$ i,k,

$$\hat{u}_{ik} = \left[\frac{1}{\sum_{j=1}^{c} \left(\frac{||x_k - \hat{v}_i||}{||x_k - \hat{v}_j||} \right)^{\frac{2}{m-1}}} \right], 1 \leq i \leq c; 1 \leq k \leq n, \quad (6a)$$

$$\hat{v}_i = \left[\frac{\sum_{k=1}^{n} (\hat{u}_{ik})^m x_k}{\sum_{k=1}^{n} (\hat{u}_{ik})^m} \right], 1 \leq i \leq c \quad (6b)$$

are necessary in order for $(\hat{U}, \hat{v})$ to be a local solution of (5). Full details for m=1 and the singular cases $x_k = \hat{v}_i$ for some i and k may be found in References 16 and 17. At m=1 requirement (6a) is replaced by a nearest neighbor assignment rule, $U \epsilon M_c$ is necessarily hard, the cluster centers in (6b) are merely the centroids of the hard subsets in U, and the resultant algorithm is essentially the hard ISODATA process of Ball and Hall.[25] For m>1 equations (6) define the

Fuzzy ISODATA algorithms

Choose any c×n matrix $U_o \epsilon M_{fc}$. (7a)

Compute the weighted means $\{\hat{v}_i\}$ with U_o and (6b). (7b)

Update $U_o \rightarrow \hat{U}$ with equation (6a). (7c)

Compute the maximum membership defect

$$\max_{i,k}\{|(\hat{U})_{ik} - (U_o)_{ik}|\}. \quad (7d)$$

If less than some prespecified tolerance ϵ, stop. Otherwise relabel $\hat{U} \rightarrow U_o$ and return to (7b).

Implicit in (7) are tie-breaking rules and resolution of singularities. These equations define an iterative optimization procedure for locating approximate minima of J_m. It is convenient to recast this loop in the form of the iterative matrix operator $T_m: M_{fc} \rightarrow M_{fc}$ defined by

$$T_m(\hat{U}_k) = \hat{U}_{k+1} = (T_m)^k(U_o), \; k = 0,1,2, \ldots . \quad (8)$$

Since U's which are part of optimal pairs for J_m must lie among the fixed points of T_m, we call approximate minima of J_m fixed points of fuzzy ISODATA. J_m has the descent property on successive iterates of T_m and the associated set of cluster centers they determine, but it is not now known whether the iterate sequence $\{T_m(\hat{U}_k)\}$ is theoretically convergent. We mention this because the numerical example below suggests an interesting conjecture about these fixed points. The possibility of using T_m to approximate a maximum likelihood operator for certain problems in unsupervised learning is discussed in Reference 26.

SCALAR MEASURES OF PARTITION QUALITY

Since optimal partitionings of X are defined as part of solutions of (5), an obvious way to rank competing partitions is by their corresponding values with J_m. Unfortunately, J_m is not an exception to the fact that global minima of objective functions may suggest very poor interpretations of substructure in X.[24,27,28] Consequently, values of J_m do *not* necessarily rank the merits of different $\hat{U}$'s in M_{fc} as worthwhile clusterings of X. It is here that fuzzy ISODATA departs from conventional clustering techniques, because with hard objective functions the functional values are the only information usually available for addressing this question. With fuzzy partitions however, the fuzziness of $\hat{U}$ allows one to associate various measures of partition quality with $\hat{U}$ which are independent of the method used to produce these partitions. Fuzzy ISODATA is used to generate likely candidates for optimal clusterings of X; their relative quality has been assessed by either of two scalar valued measures defined on M_{fc}:

$F_c(U) = \text{trace}(UU^t)/n$, superscript t being here transpose, (9)

$$H_c(U) = -\left(\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik} \log_a u_{ik}\right)/n, \text{ with } a \in (1,\infty). \quad (10)$$

$F_c: M_{fc} \rightarrow [1/c,1]$ was defined in Reference 17 as the *partition coefficient* of U; $H_c: M_{fc} \rightarrow [0,\log_a c]$ was defined in Reference 27 as the *average classification entropy* of U. Although the functional forms of F_c and H_c are quite different, they are related as follows:

$$F_c(U) = 1 \Leftrightarrow H_c(U) = 0 \Leftrightarrow U \in M_c \text{ is hard.} \quad (11a)$$

$$F_c(\tilde{U}) = 1/c \Leftrightarrow H_c(\tilde{U}) = \log_a c \Leftrightarrow \tilde{U} = [1/c]. \quad (11b)$$

$$1 - F_c(U) < \left(\frac{H_c(U)}{\log_a e}\right) < \tfrac{1}{2}(c - F_c(U)). \quad (11c)$$

Equations (11) suggest that the equi-membership partition $\tilde{U} = [1/c]$, i.e., $\tilde{u}_{ik} = 1/c \; \forall i,k$, is the fuzziest or worst one can do (geometrically $\tilde{U}$ is the centroid of M_{co}); on the other hand, the ideal situation occurs when the substructure in X is so distinct that a fuzzy algorithm recommends a hard c-partitioning of X. Maximizing F_c over different fixed points of fuzzy algorithms minimizes the total content or overlap in pairwise fuzzy intersections; equivalently, minimizing H_c over the same choices maximizes the "information" extracted from U. In either case, we presume that values of these measures serve as a relative indication of the uncertainty an *algorithm* experiences in trying to assign memberships to the vectors in X. Note that F_c and H_c are well defined for partitions generated by *any* fuzzy clustering method, not just ISODATA; moreover, these functions convey no information about the relative merits of hard c-partitions of X, their usefulness depending entirely on the idea of fuzziness. Numerical evidence indicates that H_c is probably more sensitive than F_c in ranking U's; this has been attributed to the fact that the slope of the logarithmic curve on most of (0,1) is much steeper than that of the parabola (H_c is a sum of logarithmic terms, F_c a sum of parabolic ones). Nonetheless, both measures seem useful, since the lower bound in (11c) is a sharper indication than that in (11a) of how small $H_c(U)$ is.

In general the clustering strategy used with fuzzy ISODATA has been to minimize H_c over approximate fixed points of T_m for whatever alternatives have been considered, and regard the resultant c-partitioning of X as the most optimal one. If this partition is relatively fuzzy (as measured by H_c), we do not infer that X has no well defined substructure; we conclude that none of the algorithms tried have been successful at finding it.

A NUMERICAL EXAMPLE

Table I lists 11 symptoms of 107 stomach disease patients who have either hiatal hernia (patients 1-57) or gallstones (patients 58-107). Table I constitutes our binary data set X. The data was collected as part of a larger study at the Henry Ford Hospital in Detroit by Rinaldo, Scheinok, and Rupe.[30] Various studies utilizing the larger data set for computerized medical

TABLE I—Data Set X: Class 1; Hiatal Hernia

Patient	Symptoms										
1	0	1	0	0	0	1	0	0	0	0	0
2	0	1	0	0	0	1	0	0	0	0	0
3	0	1	0	0	0	1	0	1	0	0	0
4	0	1	0	0	0	1	0	1	0	0	0
5	0	1	0	0	0	1	0	1	0	0	0
6	0	1	0	0	0	1	0	1	0	0	0
7	0	1	0	0	0	1	0	1	1	0	0
8	0	1	0	0	0	1	0	1	1	0	0
9	0	1	0	0	0	1	0	1	1	0	0
10	0	1	0	0	0	1	1	0	0	0	0
11	0	1	0	0	0	1	1	0	1	0	0
12	0	1	0	0	0	1	1	0	1	0	1
13	0	1	0	0	0	1	1	1	0	0	0
14	0	1	0	0	1	1	0	1	0	0	1
15	0	1	0	1	0	1	0	0	1	0	0
16	0	1	0	1	0	1	0	0	1	0	1
17	0	1	0	1	0	1	0	1	0	1	1
18	0	1	0	1	0	1	0	1	1	0	0
19	0	1	0	1	0	1	1	0	1	0	0
20	0	1	0	1	0	1	1	0	1	0	0
21	0	1	0	1	0	1	1	0	1	0	0
22	0	1	0	1	1	0	0	1	1	0	1
23	0	1	1	0	0	0	0	1	1	0	1
24	0	1	1	0	1	0	0	0	0	0	0
25	0	1	1	0	1	0	1	0	0	1	1
26	0	1	1	1	1	0	0	1	0	0	0
27	1	0	0	0	0	0	0	0	0	0	0
28	1	1	0	0	0	0	0	0	0	0	0
29	1	1	0	0	0	0	0	0	1	0	0
30	1	1	0	0	0	0	0	0	1	0	0
31	1	1	0	0	0	0	1	0	0	0	0
32	1	1	0	0	0	0	1	0	1	0	0
33	1	1	0	0	0	1	0	0	1	0	0
34	1	1	0	0	0	1	0	0	1	0	0
35	1	1	0	0	0	1	0	1	0	0	0
36	1	1	0	0	0	1	0	1	0	0	0
37	1	1	0	0	0	1	0	1	1	0	0
38	1	1	0	0	0	1	0	1	1	0	0
39	1	1	0	0	0	1	1	0	0	0	0
40	1	1	0	0	0	1	1	0	0	0	0
41	1	1	0	0	0	1	1	0	1	0	0
42	1	1	0	0	0	1	1	0	1	0	0
43	1	1	0	0	0	1	1	0	1	0	0
44	1	1	0	0	1	0	0	0	1	0	0
45	1	1	0	0	1	0	1	0	0	0	0
46	1	1	0	0	1	0	1	0	0	0	0
47	1	1	0	0	1	0	1	0	1	0	0
48	1	1	0	0	1	1	1	1	1	0	0
49	1	1	0	1	0	0	0	0	1	0	1
50	1	1	0	1	0	1	0	0	0	0	0
51	1	1	0	1	0	1	0	0	0	0	1
52	1	1	0	1	0	1	0	1	0	0	0
53	1	1	0	1	1	0	1	0	0	0	0
54	1	1	0	1	1	0	1	1	1	0	1
55	1	1	1	0	0	1	0	1	1	0	0
56	1	1	1	0	0	1	1	1	1	0	0
57	1	1	1	0	1	1	1	0	0	0	1

TABLE I (Continued)

Patient	Symptoms										
58	0	0	1	0	0	0	0	1	0	1	1
59	0	0	1	0	0	1	0	1	0	0	0
60	0	0	1	0	0	1	0	1	0	0	1
61	0	0	1	0	1	0	0	1	0	1	0
62	0	0	1	1	0	1	0	0	0	0	0
63	0	0	1	1	0	1	0	0	0	0	0
64	0	0	1	1	0	1	0	1	0	0	0
65	0	0	1	1	0	1	0	1	0	0	0
66	0	0	1	1	0	1	0	1	0	0	0
67	0	1	0	0	0	0	1	0	0	0	0
68	0	1	0	0	0	1	0	0	0	0	0
69	0	1	0	0	0	1	0	1	0	0	0
70	0	1	0	0	0	1	0	1	0	0	0
71	0	1	0	0	0	1	1	0	0	0	0
72	0	1	0	0	0	1	1	1	0	0	0
73	0	1	0	0	0	1	1	1	0	0	0
74	0	1	0	1	0	1	0	0	0	0	0
75	0	1	0	1	1	1	1	1	0	0	0
76	0	1	1	0	0	0	0	1	0	0	1
77	0	1	1	0	0	1	0	1	0	0	0
78	0	1	1	0	0	1	0	1	0	0	0
79	0	1	1	0	0	1	0	1	0	1	0
80	0	1	1	0	0	1	0	1	0	1	1
81	0	1	1	1	0	1	0	1	0	0	0
82	0	1	1	1	0	1	0	1	0	0	0
83	0	1	1	1	0	1	0	1	0	0	0
84	0	1	1	1	0	1	0	1	0	0	0
85	0	1	1	1	0	1	0	1	0	0	0
86	0	1	1	1	0	1	0	1	0	0	0
87	0	1	1	1	0	1	0	1	0	0	1
88	0	1	1	1	0	1	0	1	0	1	0
89	0	1	1	1	0	1	1	1	0	0	0
90	1	0	1	0	0	1	0	1	0	0	0
91	1	0	1	1	0	1	0	1	0	0	0
92	1	0	1	1	0	1	0	1	0	0	0
93	1	0	1	1	0	1	0	1	0	0	0
94	1	1	0	0	0	1	0	0	0	0	0
95	1	1	0	0	0	1	0	0	1	0	0
96	1	1	0	0	0	1	0	1	0	0	0
97	1	1	0	0	0	1	0	1	1	0	0
98	1	1	0	0	0	1	0	1	1	1	0
99	1	1	0	0	1	1	0	1	1	0	0
100	1	1	0	1	0	1	0	0	0	0	0
101	1	1	1	0	0	1	0	0	0	0	1
102	1	1	1	0	0	1	0	1	0	0	0
103	1	1	1	0	0	1	0	1	0	0	0
104	1	1	1	0	0	1	0	1	0	0	1
105	1	1	1	0	0	1	1	0	0	0	0
106	1	1	1	1	0	1	0	0	0	0	0
107	1	1	1	1	0	1	0	0	0	0	0

diagnosis have been discussed in References 31-35; full details on the data are available in Reference 31. The 11 symptoms measured (present=1, absent=0) were:

Symptom	*Description or type of Abdominal Pain*
1	Male=1; Female=0
2	Epigastric Pain
3	Upper right quadrant pain
4	Back pain
5	Discomfort episodes of 1-4 weeks
6	Discomfort episodes of 0-1 days
7	Relief induced by food ingestion
8	Aggravation induced by food ingestion
9	Aggravation induced by position
10	Weight loss (at least 20 lbs. in 6 mos.)
11	Persistence (at least 1 month in length)

The calculations were made in single precision Fortrans IV using logarithms in (10) to base e=2.718.... The convergence threshold ϵ used in (7b) was $\epsilon=0.01$. Because fuzzy ISODATA—like all hill climbing methods—is susceptible to stagnation at local minima of J_m, it is necessary to test the stability of fixed points of T_m by varying the initial guess for U_o in (7a). Other studies report results concerning this parameter; for this example the only initial guess used is

$U_o=\alpha U+\beta\tilde{U}$, where $\alpha=\sqrt{1/2}$, $\beta=1-\sqrt{1/2}$, and

$$U=\left[\begin{array}{c|c} \underbrace{\begin{matrix}1&0&\cdots&0\\0&1&\cdots&0\\ \vdots&&&\vdots\\0&0&\cdots&1\end{matrix}}_{c\times c} & \underbrace{\begin{matrix}1&1&\cdots&1\\0&0&\cdots&0\\ \vdots&&&\vdots\\0&0&\cdots&0\end{matrix}}_{c\times(n-c)} \end{array}\right] \qquad (12)$$

U_o is an initial guess lying midway between $\tilde{U}$ and the hard c-partitions of X, as measured by the value of F_c, since $F_c(U_o)=1/2+(1/2c)$. Only one initial guess is used in this example to shorten the presentation of numerical results.

Finding c is often the most important and difficult problem in clustering. The use of fuzzy ISODATA for this purpose is discussed elsewhere; in this investigation we fix c=2 in the interests of brevity. The algorithmic parameters varied here are the weighting exponent m and norm $\|\cdot\|$ appearing in (4). Values for m are 1.10, 1.33, 1.67, and 2.00. Results with other values are contained in Reference 29. Three norms induced by the weighted inner product $\langle x,x\rangle=x^tAx$ on R^d were used. These norms were realized by different choices for the symmetric matrix A:

N1 (Euclidean) induced by A=I, the d×d identity. (13a)

N2 (Diagonal) induced by $A=[\mathrm{diag}(\sigma_1^2,\ldots,\sigma_d^2)]^{-1}$, the inverse of the diagonal matrix of marginal sample feature variances. (13b)

N3 (Mahalonobis) induced by $A=[\mathrm{cov}(X)]^{-1}$, the inverse of the sample covariance matrix. (13c)

Further discussion on these choices may be found in Reference 35. Having established the computing protocols, we turn to the numeric results.

Table II lists entropies H_2 and their lower bounds $1-F_2$ for the fixed points of T_m obtained by processing X with (7) under the assumptions above. These values are comparable only for fixed values for m, because as m→1, partitions obtained by fuzzy ISODATA are always "less fuzzy" in the sense of F_c or H_c. Since (7) represents an infinite family of algorithms, there is the practical question of which one to use. The only theoretical result concerning this to date appears in Reference 36, where an analogy to minimum resistance electrical networks is used to suggest that only J_2 extends the physical interpretation of J_1 made there. It will be seen in Table II that ISODATA proceeds from U_o to $\tilde{U}$ (quite rapidly) for every norm at m=2.00; for N2 and N3 at m=1.67; and for N3 at m=1.33 (recall from (11b) that with c=2, $F_2=.500$ if and only if evaluated at $\tilde{U}$). Whether or not other initial guesses for U_o would lead to this fixed point is a matter of speculation; the rather surprising conjecture suggested by this observation is that the size of stability domains of fixed points of T_m is dependent on both m and the norm in (4). Of course, as m→∞, $\tilde{U}$ becomes the *only* fixed point of T_m, as is evident from (6a). Table II shows that it may be necessary to experimentally decrease m towards m=1 until fuzzy ISODATA successfully begins to avoid equi-memberships for a given set of data.

The values in Table II also indicate a slight preference for the Euclidean norm over N2, and a definite preference compared to N3, so we infer that this data

TABLE II—Entropies for Data Set X

Weighting Exponent m	Norm $\|\cdot\|$	Lower Bound $1-F_2(\hat{U})$	Entropy $H_2(\hat{U})$
1.10	N1	.051	.088
	N2	.057	.095
	N3	.086	.162
1.33	N1	.253	.397
	N2	.274	.425
	N3	.500	.693
1.67	N1	.420	.608
	N2	.500	.693
	N3	.500	.693
2.00	N1	.500	.693
	N2	.500	.693
	N3	.500	.693

is most separable by N1 among the three norms considered. Accordingly, the norm in (4) for subsequent runs is now fixed at $||\cdot||=N1$, and in view of the results obtained at $m=1.67$ and 2.00, we drop these values for the weighting exponent.

In Table III are listed the membership functions corresponding to terminal partitions obtained with (7), $||\cdot||=N1$, and $m=1.10$, 1.33. Scanning these values, one quickly obtains a feel for which members of the data indicate a strong desire to be classified into one subclass or the other. As we expect, the partition of X associated with the smaller value of m is very nearly hard, while the second partition begins to exhibit clearly those subjects in X apparently causing the most difficulty to ISODATA in assigning memberships. This feature of fuzziness—*identification of the troublesome or distinguished individuals in the data*—is perhaps the most important reason for using fuzzy models. Information of this kind is simply not available when using hard classification procedures, for then all the entries in solution partitions are 0's and 1's.

Since any discussion of error rates presumes a comparison with hard labels, it is necessary to convert fuzzy partitions into hard ones before this is possible. An obvious (but not necessarily best) way to do this is via the maximum membership rule: assign each $x_k \epsilon X$ to the cluster in which it holds maximum membership. All error rates mentioned below are computed with hard 2-partitions of the data obtained in this fashion. With this convention in mind, we have from Table III at $m=1.10$ 23 incorrect labels, and at $m=1.33$, 25 mislabelled patients. We emphasize that these are *not* classifier performance rates, because we are clustering here; no attempt is being made to train a classifier for prediction with unlabelled samples. However, we presume these figures are indicative of error rates which may be obtained with fuzzy classifiers now under study. Of more immediate interest is the way we can use fuzzy ISODATA to attack the feature selection problem.

FEATURE SELECTION USING FUZZY ISODATA

Contrary to one's intuition, adding more features does not always lead to better classifier performance.[3] In some instances the converse is true; deletion of features may remove the source of confusion preventing an algorithm from detecting substructure known (or presumed) to exist in the data, and in any event, reduction of the dimension of feature space alleviates the computational burden imposed by using many features. In medicine, this amounts to asking for the minimum number of symptoms needed to detect a particular disease, or to discriminate between closely related ones. The basis for a technique of fuzzy feature selection using algorithm (7) is contained in the simple

Proposition Let X be any binary valued data set, $X=\{x_1, \ldots, x_n\}$ contained in R^d; let $\{\hat{v}_i\}$ be the cluster centers given by (6b); and suppose $\hat{v}_{ij}$ to be the j^{th} component of $\hat{v}_i$ for $1\leq i\leq c$; $1\leq j\leq d$. Then

$$0\leq \hat{v}_{ij}\leq 1 \ \forall \ i,j \tag{14a}$$

$$\hat{v}_{ij}=0 \Rightarrow x_{1j}=x_{2j}=\ldots=x_{nj}=0 \tag{14b}$$

$$\hat{v}_{ij}=1 \Rightarrow x_{ij}=x_{2j}=\ldots x_{nj}=1 \tag{14c}$$

Proof Rewriting equation (6a) in the form $\hat{u}_{ik}=(1/1+c_{ik})$, where c_{ik} is the sum in the denominator over $j=1,2, \ldots, c$ with $j\neq i$, we observe that $c_{ik}>0$ for every i and k, hence $\hat{u}_{ik} \epsilon (0,1)$ for every i and k. In view of this the denominator in (6b), $\sum_{k=1}^{n} (\hat{u}_{ik})^m>0$ for $1\leq i\leq c$. Now consider the component form of (6b) for any j;

$$\hat{v}_{ij}=\sum_{k=1}^{n}\left[\frac{(\hat{u}_{ik})^m}{\sum_{s=1}^{n}(\hat{u}_{is})^m}\right]x_{kj}; \ 1\leq j\leq d; \ 1\leq i\leq c. \tag{15}$$

The coefficients of x_{kj} in (15) are all strictly positive, and since every x_{kj} is greater than or equal to zero, $\hat{v}_{ij}$ is also. Moreover, this also shows that $\hat{v}_{ij}$ can equal zero if and only if all the x_{kj} in (15) are zero. Finally, since every x_{kj} in (15) is less than or equal to one, that sum is bounded above by 1, the number obtained upon replacing all of the x_{kj}'s with 1. Since the maximum is attained when this occurs, the proof is complete.

We elaborate the implications of equations (14) for feature selection by the following series of observations:

(i) $\hat{v}_{ij}=0$: Since the proof is independent of i, it's easy to see that an even stronger statement holds: $\hat{v}_{ij}=0 \Leftrightarrow \hat{v}_{kj}=0$ for $1\leq k\leq c$ with $k\neq i$. From (14b) it follows that this occurs when and only when attribute j is absent from all n members of the data, in which case it is irrelevant to substructure in X (medically, no patient had symptom j).

(ii) $\hat{v}_{ij}=1$: As in (i), the stronger statement $\hat{v}_{ij}=1$ if and only if all the $\hat{v}_{kj}$'s with $k\neq i$ are 1 holds. In this event, feature j is a maximal descriptor of the n individuals in X (medically, all patients had symptom j).

(iii) $0<\hat{v}_{ij}<1$: Again, this can happen when and only when $0<\hat{v}_{kj}<1$ for all $k\neq i$. Ostensibly, feature j has a variable amount of influence in describing members of the c subgroups in X. This suggests that the relative magnitudes of $\hat{v}_{1j}, \hat{v}_{2j}, \ldots, \hat{v}_{cj}$ may rank the efficacy of j as a descriptor of each subclass (medically, some patients in each subclass had symptom j, and others did not).

(iv) Combining (i)-(iii), it is seen that one of the cluster centers $\hat{v}_i$ is *entirely* binary valued if and only if *all c of them are*, and this occurs if and only if all n members of the data are identical. In this eventuality there is no possibility for mathematical (or medical) detection of subclasses in X. On the other hand, we

TABLE III—Membership Functions Obtained by Fuzzy ISODATA

Patient	m=1.10		m=1.33	
	$\hat{u}_2$	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_1$
1	.001	.999	.280	.720
2	.001	.999	.280	.720
3	.137	.863	.833	.167
4	.137	.863	.833	.167
5	.137	.863	.833	.167
6	.137	.863	.833	.167
7	.002	.998	.420	.580
8	.002	.998	.420	.580
9	.002	.998	.420	.580
10	.000	1.000	.118	.882
11	.000	1.000	.048	.952
12	.000	1.000	.114	.886
13	.005	.995	.464	.536
14	.389	.611	.671	.329
15	.002	.998	.252	.748
16	.022	.978	.341	.659
17	.959	.041	.820	.180
18	.092	.908	.638	.362
19	.000	1.000	.137	.863
20	.000	1.000	.137	.863
21	.000	1.000	.137	.863
22	.233	.767	.499	.501
23	.917	.083	.685	.315
24	.735	.265	.554	.446
25	.558	.442	.466	.534
26	.999	.001	.867	.133
27	.015	.985	.230	.770
28	.000	1.000	.086	.914
29	.000	1.000	.038	.962
30	.000	1.000	.038	.962
31	.000	1.000	.049	.951
32	.000	1.000	.025	.975
33	.000	1.000	.022	.978
34	.000	1.000	.022	.978
35	.003	.997	.435	.565
36	.003	.997	.435	.565
37	.000	1.000	.152	.848
38	.000	1.000	.152	.848
39	.000	1.000	.032	.968
40	.000	1.000	.032	.968
41	.000	1.000	.013	.987
42	.000	1.000	.013	.987
43	.000	1.000	.013	.987
44	.000	1.000	.068	.932
45	.000	1.000	.079	.921
46	.000	1.000	.079	.921
47	.000	1.000	.046	.954
48	.000	1.000	.113	.887
49	.006	.994	.176	.824
50	.003	.997	.239	.761
51	.038	.962	.340	.660
52	.236	.764	.684	.316
53	.004	.996	.153	.847

TABLE III (Continued)

Patient	m=1.10 $\hat{u}_2$	m=1.10 $\hat{u}_1$	m=1.33 $\hat{u}_2$	m=1.33 $\hat{u}_1$
54	.020	.980	.258	.742
55	.507	.493	.576	.424
56	.052	.948	.331	.669
57	.149	.851	.324	.676
58	.998	.002	.834	.166
59	1.000	.000	.968	.032
60	1.000	.000	.936	.064
61	.994	.006	.796	.204
62	1.000	.000	.884	.116
63	1.000	.000	.884	.116
64	1.000	.000	.978	.022
65	1.000	.000	.978	.022
66	1.000	.000	.978	.022
67	.000	1.000	.126	.874
68	.001	.999	.280	.720
69	.137	.863	.833	.167
70	.137	.863	.833	.167
71	.000	1.000	.118	.882
72	.005	.995	.464	.536
73	.005	.995	.464	.536
74	.059	.941	.546	.454
75	.155	.845	.576	.424
76	.998	.002	.865	.135
77	1.000	.000	.987	.013
78	1.000	.000	.987	.013
79	1.000	.000	.940	.060
80	.999	.001	.904	.096
81	1.000	.000	.992	.008
82	1.000	.000	.992	.008
83	1.000	.000	.992	.008
84	1.000	.000	.992	.008
85	1.000	.000	.992	.008
86	1.000	.000	.992	.008
87	1.000	.000	.967	.033
88	1.000	.000	.960	.040
89	.999	.001	.909	.091
90	.999	.001	.883	.117
91	1.000	.000	.923	.077
92	1.000	.000	.923	.077
93	1.000	.000	.923	.077
94	.000	1.000	.068	.932
95	.000	1.000	.022	.978
96	.003	.997	.435	.565
97	.000	1.000	.152	.848
98	.005	.995	.283	.717
99	.001	.999	.191	.809
100	.003	.997	.239	.761
101	.731	.269	.525	.475
102	.997	.003	.888	.112
103	.997	.003	.888	.112
104	.995	.005	.836	.164
105	.034	.966	.239	.761
106	.958	.042	.674	.326
107	.958	.042	.674	.326

find that if and only if a single cluster center has *no* component either 0 or 1, then all c cluster centers are of this type.

In view of these remarks, it seems natural to call the components $\{\hat{v}_{ij}\}$ of cluster center $\hat{v}_i$ the *feature centers* of class i. Table IV exhibits values of these centers for each of the fuzzy partitions listed in Table III. The ranking of symptom importance for patients with hiatal hernia (class 1) established by values of $\{\hat{v}_{ij}\}$ at either value of m is $2>6>1>9 \ldots >10$. We infer from this that among the 11 symptoms measured, epigastric pain (2) is most likely to occur in patients with this disorder, whereas they will exhibit weight loss (10) only occasionally. To see whether the magnitudes of the $\hat{v}_{1j}$'s really do this, let p_{ij} be the relative frequency of occurrence of symptom j in class i patients. From Table I we find that $p_{12}=0.982$, $p_{1,10}=0.035$. These frequencies should be compared to the values of the corresponding feature centers for class 1: for example, with m=1.10 we have $\hat{v}_{12}=0.985$, and $\hat{v}_{1,10}=0.021$. These comparisons seem to corroborate our supposition concerning the ability of the fuzzy feature centers to rank the significance of the features as descriptors of each class.

For patients with gallstones (class 2), there is some shifting in ranks established by changing m from 1.10 to 1.33; this seem to indicate that members of this class are somewhat less distinctive. Nonetheless, we find from Table IV that in both cases, the most important features are {3,6,8,2}; the least important are {5,7,9}. Of course, one may take the opposite view, and regard {5,7,9} as the features most important for deciding a patient does *not* have gallstones. This remark points up the fact that the $\hat{v}_{ij}$'s do not establish which features possess discriminatory power for separating class i from closely related classes, and at the same time, suggests a way to use the feature centers for *pairs* of subclasses to select optimal discriminators.

TABLE IV—Cluster Centers for the Membership Functions in Table III

Exponent	Symptom	Feature Centers (Hernia)	Feature Centers (Galls.)	Absolute Differences
m	j	$\hat{v}_{1j}$	$\hat{v}_{2j}$	$\lvert\hat{v}_{1j}-\hat{v}_{2j}\rvert$
1.10	1	.570	.269	.302
	2	.985	.668	.317
	3	.063	.929	.865
	4	.226	.551	.324
	5	.174	.104	.070
	6	.770	.837	.068
	7	.418	.048	.370
	8	.393	.844	.451
	9	.479	.044	.435
	10	.021	.165	.144
	11	.117	.251	.134
1.33	1	.654	.260	.394
	2	.974	.752	.222
	3	.105	.686	.581
	4	.214	.485	.271
	5	.191	.098	.093
	6	.713	.878	.164
	7	.467	.092	.375
	8	.285	.839	.553
	9	.527	.103	.423
	10	.031	.118	.087
	11	.127	.198	.071

An obvious indication of "how separable" classes i and j are is their cluster center separation $||\hat{v}_i-\hat{v}_j||$. This measure, however, suppresses the information we want to use for reducing the number of features required to effect the classification. A more suitable measure is afforded by the vector of absolute differences of the components of $\hat{v}_i$ and $\hat{v}_j$: for all i and j let

$$f_{ij}=(|\hat{v}_{i1}-\hat{v}_{j1}|,|\hat{v}_{i2}-\hat{v}_{j2}|,\ldots,|\hat{v}_{id}-\hat{v}_{jd}|). \quad (16)$$

The components of $f_{ij,k}$ of vector f_{ij} measure feature center separations between the feature centers for classes i and j. Equations (14) lead to the following results for these components:

$$0\leq f_{ij,k}\leq 1 \text{ for all i,j, and k.} \quad (17a)$$

$$f_{ij,k}=0 \Leftrightarrow \text{Either all or none of the vectors in both classes i and j have feature k.} \quad (17b)$$

$$f_{ij,k}=1 \Leftrightarrow \text{All vectors in class i and no vectors in class j have feature k, or vice versa.} \quad (17c)$$

We presume feature k to be either useless or optimal as a discriminator between classes i and j according as (17b) or (17c) respectively occurs. (17a) shows these to be the extremes, intimating that the values $f_{ij,1}$, $f_{ij,2}, \ldots, f_{ij,d}$ rank by their magnitudes the relative utility of the d features for discrimination between classes i and j.

To test this speculation, the vector f_{12} defined by (16) corresponding to the cluster centers in Table IV was used to identify the optimal feature subsets of dimensions 1,2, and 3, and the data set X was reprocessed with fuzzy ISODATA using only these features. The last column of Table IV reports the values of $f_{12,k}$: evidently symptom 3—upper right quadrant pain—is implicated as the most powerful attribute for distinguishing between gallstones and hiatal hernia. The feature center values $\hat{v}_{13}=0.105$ and $v_{23}=0.686$ suggest that very few hernia patients suffer from symptom 3, while most gallstones patients may be expected to have it. Indeed, from Table I we find that the relative frequencies of symptom 3 are $p_{13}=0.123$ and $p_{23}=0.680$ respectively. Continuing in this fashion, we deduce that either {3,8} or {3,9} would be the best 2-dimensional subset of features to use; that {3,8,9} is the best set of 3 features at either value of m; and so on.

The results of clustering these feature subsets are reported in Table V as numbers of misclassifications stemming from the hard 2-partitions realized by maximum membership conversion of the associated fuzzy fixed points of T_m. Using symptom 3 alone results in exactly the same hard partitions as using symptoms 3 and 9; moreover, it will be seen that the overall error rates achieved with either of these subsets is at least

TABLE V—Misclassifications * Using Reduced Feature Spaces

Symptoms Used	m=1.10 Galls. n_1=50	Hernia n_2=57	Overall n=107	m=1.33 Galls. n_1=50	Hernia n_2=57	Overall n=107
1-11	17	6	23	13	12	25
3	16	7	23	16	7	23
3,9	16	7	23	16	7	23
3,8	13	23	36	13	23	36
3,8,9	13	23	36	10	17	27
Deleted#	n_1=43	n_2=41	n=84	n_1=43	n_2=41	n=84
1-11	0	7	7	0	7	7

*Based on hard (maximum membership) partitions.

#Data X with patients {23-26,55-57,67-75,94-100} deleted.

as good as the rate attained using all 11 features. Note that symptoms 3 and 9 are much less effective than 3 and 8, and the error rate using {3,8,9} is in between the best and worst ones shown. From these results it appears that the feature selection method proposed above successfully extracts a small number of features which possess essentially the same information relevant to substructure in X detected by fuzzy ISODATA as the original ones.

SUMMARY

Fuzzy clustering, and in particular fuzzy ISODATA, has been reviewed, and is proposed here as a basis for a new technique applicable to the problem of feature selection. Specifically, equations (14) and (17) seem useful in ranking the effectiveness of binary valued features both as subclass representatives and as discriminators between pairs of fuzzy subclasses in X. A numerical example was presented which seems successful enough to warrant further investigations into the plausibility of the method. We note that this technique is applicable *only* for binary data sets: in fact, (6b) shows that the cluster centers $\{\hat{v}_i\}$ lie in the linear subspace generated by the data, so when the features have continuous domains, (14) and (17) are invalid.

As a means of computerized medical diagnosis, the technique described above is incomplete in the sense that it is a clustering method, not a classifier. Nonetheless, it seems fair to assert that our example exhibits the promise fuzzy sets may hold for this problem. Our conviction is that fuzziness is the premise needed as a basis for pattern recognition; more precisely, we think it an appropriate generalization of the conventional strategies criticized in Reference 5. The reason for this lies with the fuzzy membership values generated by algorithms like fuzzy ISODATA; not only do they indicate a patient's relative affinity for having every disease represented by members of the data; but perhaps more importantly, low memberships can be used to identify those patients whose symptoms indicate further personal attention. For example, the values in Table III suggest that the 23 patients whose Table I labels are {23-26,55-57,67-75,94-100} are—by virtue of their relatively low memberships—the ones most affecting the computer's success at separating the two subclasses. If these 23 individuals are deleted from X, and the remaining 84 patients are processed with ISODATA, the results reported in the last row of Table V indicate an increase of about 14 percent in the accuracy of labelling obtained on all 11 features with either value of m. This appears to confirm that low memberships signal troublesome patients. (Note that processing this deleted set with only feature 3 results in a recognition rate of 100 percent: the 23 patients identified above are precisely the 23 subjects having the "uncharacteristic" labels for members of their classes with respect to feature 3 alone). Of course, it is not the business of the medical community to *delete* troublesome patients from data sets for the convenience of a computer: on the contrary, these are the patients that doctors want most to *identify*, and we believe that fuzzy methodologies will eventually be useful in realizing computer assistance and counseling for people in this profession.

REFERENCES

1. Zadeh, L., "Fuzzy Sets," *Inf. and Control*, 8, 1965, pp. 338-353.
2. Zadeh, L., K. Fu, K. Tanaka and M. Shimura, *Fuzzy Sets and Their Applications to Cognitive and Decision Processes*, Academic Press, New York, 1975.
3. Duda, R. and P. Hart, *Pattern Classification and Scene Analysis*, Wiley-Interscience, New York, 1973.
4. Tou, J. and R. Gonzalez, *Pattern Recognition Principles*, Addison Wesley, Reading, 1975.
5. Croft, D., "Mathematical Methods in Medical Diagnosis," *Annals Bio. Engr.*, 2, 1974, pp. 68-89.
6. Bellman, R., R. Kalaba, and L. Zadeh, "Abstraction and Pattern Classification," *Jo. Math. Anal. and Appl.*, 13, 1966, pp. 1-7.
7. Chang, C., *Fuzzy Sets and Pattern Recognition*, PhD Thesis, Univ. of California at Berkeley, 1967.
8. Wee, W., *On a Generalization of Adaptive Algorithms and Applications of the Fuzzy Set Concept to Pattern Classification*, Tech. Rep. 67-7, EE Dept., Purdue Univ., Lafayette, Indiana, 1967.
9. Flake, R. and B. Turner, "Numerical Classification for Taxonomic Problems," *Jo. Theo. Bio.*, 20, 1968, pp. 260-270.
10. Gitman, I. and M. Levine, "An Algorithm for Detecting Unimodel Fuzzy Sets and Its Application as a Clustering Technique," *IEEE Trans. Comp.*, C-19, 1970, pp. 917-923.
11. Ruspini, E., "A New Approach to Clustering," *Inf. and Control*, 15, 1969, pp. 22-32.
12. ———, "Numerical Methods for Fuzzy Clustering," *Inf. Sciences*, 2, 1970, pp. 319-350.
13. ———, *New Experimental Results in Fuzzy Clustering*, Internal Report, Brain Research Institute, UCLA, 1970.

14. ———, *Applications of Fuzzy Clustering to Pattern Recognition,* Internal Report, Brain Research Institute, UCLA, 1970.
15. Larsen, L., E. Ruspini, J. McNew, D. Walter and W. Adey, "A Test of Sleep Staging Systems in the Unrestrained Chimpanzee," *Brain Research,* 40, 1972, pp. 319-343.
16. Dunn, J., "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Jo. Cybernetics,* 3,3, 1974, pp. 32-57.
17. Bezdek, J., *Fuzzy Mathematics in Pattern Classification,* PhD Thesis, Applied Mathematics, Cornell Univ. Ithaca, 1973.
18. Zadeh, L., "Similarity Relations and Fuzzy Orderings," *Inf. Sciences,* 3, 1971, pp. 177-200.
19. Tamura, S., S. Niguchi and K. Tanaka, "Pattern Classification Based on Fuzzy Relations," *IEEE Trans.* SMC, SMC-1, 1971, pp. 61-66.
20. Dunn, J., "A Graph-Theoretic Analysis of Pattern Classification via Tamura's Fuzzy Relation," *IEEE Trans.* SMC, SMC-4, 1974, pp. 310-313.
21. Kandel, A. and L. Yelowitz, "Fuzzy Chains," *IEEE Trans.* SMC, SMC-4, pp. 472-475.
22. Yeh, R. and S. Bang, *Fuzzy Relations, Fuzzy Graphs, and their Applications to Clustering Analysis,* CS Report SESLTC-3, Univ. of Texas at Austin, 1974.
23. Rosenfeld, A., "Fuzzy Graphs," in *Fuzzy Sets and Their Applications to Cognitive and Decision Processes,* Zadeh et al., eds., Academic Press, New York, 1975.
24. Bezdek, J., "Cluster Validity with Fuzzy Sets," *Jo. Cybernetics,"* 3,3, 1974, pp. 58-73.
25. Ball, G. and D. Hall, "A Clustering Technique for Summarizing Multivariate Data," *Behav. Sci.,* 12, 1967, pp. 153-155.
26. Bezdek, J. and J. Dunn, "Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Distributions," *IEEE Trans. Comp.,* Aug., 1975, pp. 835-838.
27. Wishart, D., "Mode Analysis: A Generalization of Nearest Neighbor Which Reduces Chaining Effects," in *Numerical Taxonomy,* Cole, A. ed., Academic Press, New York, 1969, pp. 282-308.
28. Ling, R., *Cluster Analysis,* PhD Thesis, Yale Univ., New Haven, 1971.
29. Bezdek, J., "Mathematical Models for Systematics and Taxonomy," in *Proc. Eighth Int. Conf. on Numerical Taxonomy,* G. Estabrook, Ed., Freeman, San Francisco, 1975.
30. Rinaldo, J., P. Scheinok, and C. Rupe, "Symptom Diagnosis: A Mathematical Analysis of Epigastric Pain," *Ann. Int. Medicine,* 59, 1963, pp. 145-154.
31. Scheinok, P. and J. Rinaldo, "Symptom Diagnosis: Optimal Subsets for Upper Abdominal Pain," *Comp. and Bio. Res.,* 1, 1967, pp. 221-236.
32. Scheinok, P., "Symptom Diagnosis: Bayes' Theorem and Bahadur's Distribution," *Bio-Med. Comp.,* 3,1, 1973, pp. 17-28.
33. Cumberbatch, J. and H. Heaps, "Application of a Non-Bayesian Approach to Computer Aided Diagnosis for Upper Abdominal Pain," *Bio-Med. Comp.,* 1973.
34. Toussaint, G. and P. Sharpe, "An Efficient Method for Estimating the Probability of Misclassification Applied to a Problem in Medical Diagnosis," *Comp. Bio. Med.,* 4, 1975, pp. 269-278.
35. Bezdek, J. "Numerical Taxonomy with Fuzzy Sets," *Jo. Math. Bio.,* 1,1, 1973, pp. 57-71.
36. ———, "A Physical Interpretation of Fuzzy ISODATA," *IEEE Trans.* SMC, in press.