INFORMATION AND CONTROL 15, 22–32 (1969)

A New Approach to Clustering

ENRIQUE H. RUSPINI

Space Biology Laboratory, University of California, Los Angeles

A general formulation of data reduction and clustering processes is proposed. These procedures are regarded as mappings or transformations of the original space onto a "representation" or "code" space subjected to some constraints. Current clustering methods, as well as three other data reduction techniques, are specified within the framework of this formulation. A new method of representation of the reduced data, based on the idea of "fuzzy sets," is proposed to avoid some of the problems of current clustering procedures and to provide better insight into the structure of the original data.

LIST OF SYMBOLS

R:	the real line
R^+:	the non-negative real numbers
R^n:	Euclidean n-dimensional space
X:	data set
$P(x)$:	density function defined in X
g:	grouping function
C:	representation set
K:	set of subindexes of constraints
T_k, V_k:	constraints on g
I:	optimality conditions
J:	relaxed constraints
δ:	distance function
λ:	Lagrange multiplier
$\mathbf{N}$:	natural numbers
ρ:	Watanabe's inter-relation function
S_j:	fuzzy sets
$P(S_j/x)$:	degree of belongingness of x to S_j
$P(S_j)$:	size of S_j or a priori probability of S_j
$P(x/S_j)$:	density function of the cluster S_j

M_j, M:	mean distances with respect to class j and the whole population respectively
H_L:	normalized local entropy
N:	number of fuzzy sets
W_k:	weights on general constraint
U:	threshold in clustering

1. INTRODUCTION

The purpose of this paper is to introduce a general formulation of data reduction techniques and to outline a new view of the problem of numerical classification or clustering. The object of cluster analysis is to classify experimental data in a certain number of sets where the elements of each set should be as similar as possible and dissimilar from those of other sets. This implies the existence of a measure of distance or similarity between the elements to be classified. The number of such classes may be fixed beforehand or may be a consequence of some constraints imposed on them.

Present clustering techniques are in a primitive stage of development and no known procedure is exempt from the difficulties detailed in Nagy (1968).

The problem of misclassifications originated by "bridges" or strays between sets, and that of pairs classified in different classes having a greater similarity than some pairs within the same set, represent two frequent difficulties. A survey of the methods and techniques of clustering may be found in Ball (1965).

To allow a comparative study of data reduction techniques, we present first a mathematical formulation of these processes in Section 2. Section 3 is dedicated to the characterization of some of the present techniques in terms of such formulation. This will serve to emphasize the differences between them. Our approach is introduced in Section 4.

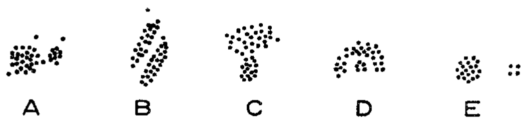

FIG. 1. Nagy (1968) has illustrated the major difficulties found in cluster analysis: (a) and (c) bridges between clusters; (b) nonspherical clusters; (d) linearly nonseparable clusters; (e) unequal cluster populations.

2. A MATHEMATICAL FORMULATION OF DATA REDUCTION TECHNIQUES

Given a set X and a probability measure P in X, then a *data reduction technique* is a set C called the *code or representation space* and a function $g:X \to C$ called the *grouping function* that satisfies the following constraint:

For each element k of a set K there are two functions,

$$\begin{aligned} T_k&:X \times X \to R, \\ V_k&:C \times C \to R, \end{aligned} \tag{1}$$

and for all $(x, y) \in X \times X$,

$$T_k(x, y) = V_k(g(x), g(y)). \tag{2}$$

In other words, we are trying to classify a "sample" of X characterized by the density function P, and C is our set of names or representatives for our original data contained in X. The g is the function that pairs each element with its name or representation. $\{T_k, V_k\}$ is a set of constraints that requires that some properties of the elements in X should be preserved after transformation. For example, in cluster analysis "close elements should be assigned to the same class."

The set $g(X) = \{g(x):x \in X\}$ is called the *reduction* of X.

Sometimes we look not only for grouping functions that satisfy the constraints but for one that is best in some sense:

$$I:C \to R, I(g) = \text{min}. \tag{3}$$

In other instances, the constraints are so restrictive that no grouping function exists. We can relax then some of those conditions so that they will hold approximately. For example, if we have $K = \{1, \cdots, N\}$, and we decide to relax all conditions, we can make

$$\begin{aligned} J(g) &= \sum_{k=1}^{N} W_k^2 \iint [T_k(x, y) - V_k(g(x), g(y))] dP(x)\, dP(y) \\ &= \text{min}. \end{aligned} \tag{4}$$

The more general formulation for this K will be,

$$\begin{aligned} &J \subset K, \\ T_k(x, y) &= V_k(g(x), g(y)), \qquad k \in J, \\ I(g) + \lambda J(g) &= \text{min}; \qquad W_k = 0, \quad \text{if} \quad k \in J. \end{aligned} \tag{5}$$

3. SOME SPECIFIC EXAMPLES

A. Current Clustering Methods

Here X = finite sample from R^n,
P = point probability in X,
C = $\mathbf{N}$ or a subset of it,
$g(x_i)$ = m (the number of the cluster to which x_i belongs).

The properties T_k and V_k vary greatly for each particular method but they are usually stated in terms of a distance function $\delta : X \times X \to R^+$ (the non-negative real numbers) that satisfies:

$$1.\ \delta(x, x) = 0, \qquad 2.\ \delta(x, y) = \delta(y, x).$$

A typical set of constraints can be described by

$$T = \begin{cases} 1 & \text{if } \delta(x, y) \geqq U \\ 0 & \text{otherwise,} \end{cases}$$

$$V = 0 \quad \text{if } g(x) = g(y); \qquad 1 \text{ otherwise,}$$

thus requiring that points distant more than some threshold U to be placed in different clusters. A similar set may require that points sufficiently close may be classified in the same set. Constraints of the number of elements of $g(X)$ are usual. In other cases, restricting properties are not invoked and the clumping properties of g are obtained from the optimality condition.

Usually, all the constraints are not satisfied and some sort of compromise is reached. This results in the anomalies described in Section1. Even when g exists, its determination is by no means easy; reallocation and correction processes are common (*cf.* Lance and Williams, 1967; Ball, 1965). These procedures may be used sequentially to generate hierarchical classifications.

B. Factor Analysis and Related Techniques

Factor analysis is a procedure to find linear relations between data sampled from a linear space. We shall restrict our description to finite dimensional Euclidean spaces. Let, then

X = subset of R^n,
C = R^m, $\quad m < n$.

The idea is to represent $x \in X$ as

$$x = g_1 V_1 + \cdots + g_m V_m, \quad \text{where} \quad V_i \in R^n, g_i \in R, 1 \leqq i \leqq m,$$

then

$$g : x \to (g_1, \cdots, g_m),$$

and here

$$K = R^2,$$

and

$$T_{\lambda\mu}(x, y) = \lambda x + \mu y,$$
$$V_{\lambda\mu}(u, v) = \lambda u + \mu v.$$

In other words, g is required to be linear. The constraints cannot be satisfied in general so that the relaxed forms

$$\int (x - g_1(x) V_1)^2 dP(x) = \min_{g_1(x), V_1}$$
$$\int [(x - g_1(x) V_1) - g_2(x) V_2]^2 dP(x) = \min_{g_2(x), V_2}$$

and so forth, are used.

C. The Information Theoretical Correlation Method of Watanabe

X = subset of the set of all m-tuples with binary components (0 or 1); m is finite.

$C = \mathbf{N}$.

$P(x)$ is the probability function of the elements of X. A function ρ (the inter-relation) is then defined (Watanabe, 1965). This inter-relation function measures the resemblance between the two sets, determined by the number of their similar or correlated binary components. It can be shown that the inter-relation between two sets is zero if and only if they are statistically independent.

Two points x and y are said to be *stuck* if no set Y exists such that $x \in X - Y$, $y \in Y$, and

$$\rho(X - Y, Y) = \min_{Z \subset X} \rho(X - Z, Z),$$

where the minimization is carried over all nonvoid proper subsets of X.

$$\begin{aligned} T(x, y) &= 1 \quad \text{if and only if } x \text{ and } y \text{ are } \textit{stuck}, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

V is such that

$$V(u, v) = \begin{cases} 0 & \text{if} \quad u \neq v, \\ 1 & \text{if} \quad u = v. \end{cases}$$

g is selected so that Card $g(X)$ = number of classes, is a maximum. (This is indeed a consequence of the preceding constraints and as a constraint is unnecessary.)

In this method the internal structure of each subclass determines the classification in groups that are not inter-related (or "redundant"). Watanabe (1965) gives an example using an information measure as his inter-relation. Distances are not used; rather sharing of a number of attributes by each group is the reason for the common label applied to its members. The method is then applied iteratively to produce a hierarchy.

D. The Information Measure of Wallace and Boulton

Here, as in the preceding examples, the space X is a subset of R^n and C is a set of real vectors such that

(a) the first element of $g(x) \in C$ represents the number of the class to which x belongs,

(b) the next element of $g(x)$ represents the type of class to which x belongs. This can be anyone of a predefined dictionary of classes (Ex: normal, uniform, etc.),

(c) the next r elements where r is variable are the parameters of the class to which x belongs,

(d) the next s elements are devoted to specify the position of x with respect to the class (ex.: distance of x to the mean in some units in normal distributions).

The length of the vector $g(x)$ is then variable depending on the class selected and the amount of information needed to specify that class and the position of x within that set. All the attributes are used for classification while the number of possible choices for classes is limited by the size of the dictionary of classes.

Clustering is then made trying to optimize the length of the codes, subjected to some restrictions, to insure a proper encoding and identification of each data point. An information measure is used and the analysis resembles the determination of optimal lengths of codes, as done in coding theory (here, more frequent classes have longer codes than unfrequent classes and are better specified than those).

For more details, see Wallace and Boulton (1968).

4. NEW APPROACH

A. Number of Classes Fixed

The set X is usually a finite subset of R^n; then let

$$X = \{x_1, \cdots, x_m\} \subset R^n, \text{ and}$$

$$C = \{\text{set of all } N\text{-tuples with positive elements and sum of its elements} = 1\}.$$

A distance δ as described in 3A is presumed to be defined in $X \times X$ and

$$g:x \rightarrow (P(S_1/x), \cdots, P(S_N/x)). \tag{6}$$

The use of the probability notation will become clearer later.

We regard numerical classification as the process of assigning to each data point a certain degree of belongingness to each class $S_1, \cdots, S_N$. In that sense, the S_j's are "fuzzy sets" in the sense of Zadeh (1966). However, our fuzzy sets have a strong probabilistic meaning and our rules of operation are not those proposed by Zadeh but those that come naturally from probability theory. Figure 2 is an idealized example of such a classification applied to a dichotomy of a simple set. Several advantages can be noted over conventional clustering representation: Points in the "core" or center of some class will have a degree = 1 of

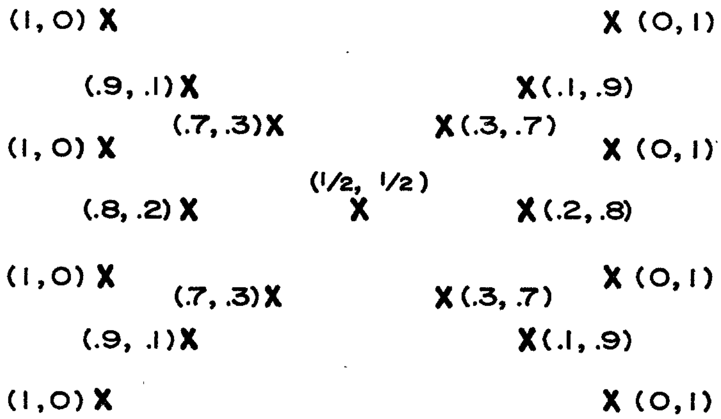

Fig. 2. A possibly "fuzzy set" dichotomy of a very simple set. The outer columns form the "core" of the set while the degree of indeterminacy varies with the distance to the core. The first number between parenthesis is the degree of belongingness to Set I and the second to Set II.

being members of that class, while boundary points, between the core and other classes may be identified as such. "Bridges" or stray points may be classified as undetermined points with a degree of indeterminacy proportional to their similarity to "core" points.

In this subsection, the number N of "fuzzy sets" is fixed. In subsection 4B, some thought is given to the problem of a variable number of subsets. Assuming now that X is finite, we introduce the following notations that will be useful later

$$P(S_j) = \sum_{i=1}^{m} P(x_i)P(S_j/x_i). \tag{7}$$

Note that $P(S_j) \geqq 0$ and $\sum_{j=1}^{N} P(S_j) = 1$. $P(S_j)$ is a measure of the relative size of each class.

$$P(x_i/S_j) = \frac{P(S_j/x_i)P(x_i)}{P(S_j)}. \tag{8}$$

From (8) it is easy to see that

$$P(x) = \sum_{j=1}^{N} P(S_j)P(x/S_j),$$

and so the clustering process may be seen as the decomposition of the density function $P(x)$ into the weighted sum of the component cluster densities $P(x/S_j)$ with weights $P(S_j)$ (the "a priori" probability of S_j). This is the main reason for the use of the probabilistic notation. The consideration of the degrees of belongingness as probabilities is very useful for pattern recognition purposes after the fuzzy sets or classes have been established.

We now define

$$M(x_i) = \sum_{k=1}^{m} P(x_k)\delta(x_i, x_k), \tag{9}$$

(mean density of the population around x_i)

$$M_j(x_i) = \sum_{k=1}^{m} P(x_k/S_j)\delta(x_i, x_k). \tag{10}$$

(mean density of the class S_j around x_i).

From (7) we see that

$$\sum_{j=1}^{N} P(S_j)M_j(x_i) = M(x_i).$$

To insure that the density functions $P(x/S_j)$ really represent clusters, optimality conditions and constraints should be imposed.

The use of both is greatly limited by computational difficulties. The procedures to carry out the numerical optimization are sophisticated enough to be, along with our numerical experiments, the object of a forthcoming paper.

Two optimality conditions have been used so far:

(I.) For each point i, let $S_{(i)}$ be such that

$$P(S_{(i)}/x_i) = \max_{1 \leqq j \leqq N} \mathrm{P}(S_j/x_i).$$

The parenthetical subscript notation is then introduced to indicate the most likely cluster assignment. Then g is selected to make

$$\sum_{i=1}^{m} \left[P(x_i) \, \frac{P(S_{(i)})M_{(i)}(x_i)}{M(x_i)} \right]^2, \tag{11}$$

a minimum.

(II.) Here g is selected to make

$$\sum_{i=1}^{m} P(x_i) \sum_{j=1}^{N} P(S_j) \left[\frac{P(S_j)M_j(x_i)}{M(x_i)} \right] P(S_j/x_i), \tag{12}$$

a minimum.

Constraint (11) tries to minimize the average mean density of a point to the cluster to which it most likely belongs. Formula (12) tries to minimize (in the average) the products $P(S_j/X_i)M_j(x_i)$ so that large mean densities will correspond to small degrees of belongingness and vice versa.

These two formulas have been applied to several examples and although some points are classified as undetermined, most are classified absolutely and the usual problems remain. For that reason, the following constraint is introduced:

$$T(x, y) = F(\delta(x, y)), \tag{13}$$

$$V(g(x), g(y)) = \left(\sum_{j=1}^{n} (P(S_j/x) - P(S_j/y))^2 \right)^{1/2}. \tag{14}$$

where $F:R^+ \rightarrow R^+$ is a nonnegative non-decreasing function not identically zero such that $F(0) = 0$. This constraint requires that the map g preserves the distance structure, transforming pairs of similar points into pairs of similar codes.

In general, this constraint cannot be satisfied by any g, and so the approach detailed in (4) is used. Here:

$$J(g) = \sum_{k=1}^{n} \sum_{i=1}^{n} P(x_i)P(x_k)(F(\delta(x_i, x_k)) - V(g(x_i), g(x_k)))^2. \quad (15)$$

Experiments with simple sets are encouraging and a computational technique is being developed. Numerical experience shows that the results of using constraints (11) and (12) provide a good starting point for the numerical methods used in the minimization of condition (15). The possibility of making F variable over a set of feasible functions to improve J is foreseen.

B. On the Number of Clusters

Some of the already defined quantities may be useful in determining the optimal number of clusters N.

First it may be noted that $J(g)$ as defined in the preceding paragraph is not a decreasing function of the number of clusters and is likely that for some N it may be a minimum.

Considerations on the diameter of a set and the distance between sets are invoked usually as constraints to determine N.

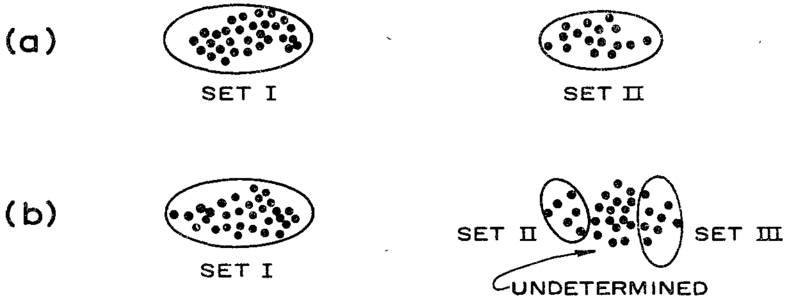

FIG. 3. The dichotomy of (a) doesn't have undetermined points and the local entropy (14) is low. The partition in three sets creates a number of undetermined points suggesting that two sets were sufficient.

In our approach, a function may be useful

$$H_L = -(\log N)^{-1} \sum_{i=1}^{N} P(x_i)\left(\sum_{j=1}^{N} P(S_j/x_i) \log P(S_j/x_i)\right) \quad (16)$$

(normalized local entropy of the classification). In the ideal case depicted in Fig. 3, the local entropy of the classification in three sets is much higher than that of two sets, corresponding to a high number of undetermined points produced as a result of the dichotomy of one cluster.

5. SUMMARY

We have presented here a mathematical formulation that permits a unified view of data reduction techniques. We have also proposed a new approach to numerical classification problems. This approach is intrinsically free of the shape and size problems of other clustering methods and using proper constraints gives better information about the structure of the data set.

Furthermore, formulation of constraints and optimality conditions is easier in terms of this framework than in the usual set description.

ACKNOWLEDGMENTS

The author wishes to thank D. O. Walter and R. T. Kado for helpful suggestions during the development of this research; M. Maddex for the typing of the manuscript; and Mrs. J. Payne for the illustrations of the article. Computational assistance was obtained from the Health Sciences Computing Facility, UCLA, supported by NIH Grant FR-3. This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by U.S. Army Medical Research and Development Command under Contract No. DADA 17-67-C-7124.

RECEIVED: February 5, 1969; revised May 28, 1969.

REFERENCES

BALL, G. H. (1965), Data analysis in the social sciences: What about the details?, Proceedings, Fall Joint Computer Conference, 533–559.

LANCE, C. N. AND WILLIAMS, W. T. (1967), A general theory of classificatory sorting strategies. II. Clustering systems, *Computer Journal* **10**, 271–277.

NAGY, G. (1968), State of the art in pattern recognition, *Proc. IEEE.* **56**, 836–882.

WALLACE, C. S. AND BOULTON, D. M. (1968), An information measure for classification, *Computer Journal* **11**, 185–194.

WATANABE, S. (1965), Une explication mathematique du classement d'objets. *In* "Information and Prediction in Science" (Dockx, S. and Bernays, P., eds.). Academic Press, New York.

ZADEH, L. A., (1965), Fuzzy sets, *Inform. Control* **8**, 338–353.