METRIC CONSIDERATIONS IN CLUSTER ANALYSIS

HERMAN CHERNOFF
STANFORD UNIVERSITY

1. Introduction and summary

A variation of the "k means" method of cluster analysis is described which is designed to take into account and profit from the possibility that the separate clusters resemble samples from multivariate normal distributions with substantially different covariance structures. This is preceded by a brief description of a standard version of the method. Indications are given when metric considerations can play an important role and a suitably modified version of the standard method is presented.

While the new method has not yet been applied it is anticipated that its most useful applications will be to situations where the clusters tend to be concentrated in nonparallel hyperplanes of the space of observations. The dimensionality of this space should not be very large. The method should require substantial sample sizes to make the implicit estimates of the covariance matrices useful.

One may expect metric considerations also to be useful in modifying other cluster analysis techniques.

2. The standard k means method

In this section we describe the k means method in the spirit of MacQueen [2]. Suppose that p represents the probability distribution of a random variable (r.v.) Z in an r dimensional Euclidean space and $|y - z|$ represents the distance between points y and z of this space. Let $S = (S_1, S_2, \cdots, S_k)$ be a *decomposition* of the space into k pairwise disjoint measurable subsets (classes) and let $x = (x_1, x_2, \cdots, x_k)$ represent k *reference* points in the space. Then

$$R(x, S) = \sum_{i=1}^{k} \int_{S_i} |z - x_i|^2 \, dp(z) \tag{2.1}$$

is a measure of the corresponding within class variance. From one point of view of the notion of cluster it would be expected that if the probability measure p corresponds to k *natural clusters*, these clusters would relate in a simple way to an (x, S) which minimizes R.

For given S, $R(x, S)$ can be minimized by selecting the reference points to be the centers of gravity, that is, $x = u(S) = (u_1(S), u_2(S), \cdots, u_k(S))$, where

$$u_i(S) = \int_{S_i} z\, dp(z)/p(S_i), \qquad i = 1, 2, \cdots, k \tag{2.2}$$

assuming $p(S_i) > 0$, $i = 1, 2, \cdots, k$. Then we have the measure

$$V(S) = R[u(S), S] = \sum_{i=1}^{k} \int_{S_i} |z - u_i(S)|^2\, dp(z). \tag{2.3}$$

Alternatively, for given reference points x, we can minimize R with respect to S by selecting a decomposition $S = T(x) = (T_1(x), \cdots, T_k(x))$ which assigns to T_i those points which are closest to x_i, that is,

$$\text{if} \quad z \in T_i(x) \quad \text{then} \quad |z - x_i| = \min_j |z - x_j|. \tag{2.4}$$

This gives us the within class variance

$$W(x) = R[x, T(x)] = \sum_{i=1}^{k} \int_{T_i(x)} |z - x_i|^2\, dp(z). \tag{2.5}$$

Thus $V(S)$ and $W(x)$ are similar but somewhat different measures of within class variance. They do coincide in the case where $T(x) = S$ and $u(S) = x$. Then $u[T(x)] = x$ and the reference points x are said to be *unbiased*.

The above described minimization properties imply that

$$V(S) = R[u(S), S] \geqq W[u(S)] \tag{2.6}$$

and

$$W(x) = R[x, T(x)] \geqq V[T(x)]. \tag{2.7}$$

Hence, given an arbitrary x, the iteration $x^1 = x, S^1 = T(x^1), \cdots, x^n = u[S^{n-1}]$, $S^n = T(x^n), \cdots$, yields the decreasing sequence

$$W(x^1) \geqq V(S^1) \geqq W(x^2) \geqq V(S^2) \geqq \cdots, \tag{2.8}$$

and hopefully converges to a pair (x, S) with a low within class variance and unbiased x. MacQueen attributes consideration of this procedure to Forgy [1] and Jennrich.

The k means method to be described may be motivated by considerations such as given above and by the aim of reducing computational labor and information storage requirements. Suppose that in place of the probability distribution p one is given a sample of independent observations on the r.v. Z. Described informally, the k means method is an iterative method of generating a sequence of reference points $x^n = (x_1^n, x_2^n, \cdots, x_k^n)$ where x^1 consists of the first k distinct observations $Z_1, Z_2, \cdots, Z_k$ on Z. Afterwards each new observation is assigned to the closest reference point which is then modified to be the average of all observations assigned to it.

More precisely let $x^1 = (Z_1, Z_2, \cdots, Z_k)$ and $w^1 = (1, 1, \cdots, 1)$. If we observe Z after x^n and $w^n = (w_1^n, w_2^n, \cdots, w_k^n)$ are formed, let

$$x_i^{n+1} = x_i^n, \quad w_i^{n+1} = w_i^n, \qquad \text{if} \quad Z \notin T_i(x^n) \tag{2.9}$$

and

$$x_i^{n+1} = \frac{w_i^n x_i^n + Z}{w_i^n + 1}, \quad w_i^{n+1} = w_i^n + 1, \qquad \text{if} \quad Z \in T_i(x^n). \tag{2.10}$$

The *weight* w_i^n is the number of observations whose mean is represented by the ith reference point x_i^n. We shall loosely refer to $T_i(x^n)$ as the ith cluster at the nth stage.

Compared with many other techniques of cluster analysis, the iterative procedure of the k means method seems to be rather economical in storage and computational requirements. At the nth stage one needs to store x^n, w^n, and the latest observation Z. The computation consists mainly of evaluating the k distances $|Z - x_i^n|$, $i = 1, 2, \cdots, k$. A sound comparison of computational efficiency would require some insight into the number of iterations required with this technique. Many alternative approaches seem to require the computation and storage of ${}_mC_2$ distances to study a sample of m points. When m is large this may be excessive. However it is possible that sampling techniques may be applied to reduce these requirements.

MacQueen [3] has proved two theorems listed below which indicate that the $W(x^n)$ of the k means method converge to $W(x)$ for an unbiased x and that the x^n converges to $u(S(x^n))$ in a weak sense. These theorems are proved under the assumptions: (i) p is absolutely continuous with respect to Lebesgue measure, and (ii) $p(R) = 1$ for a closed and bounded convex set R and $p(A) > 0$ for every open set $A \subset R$.

THEOREM 1. *The sequence of random variables* $W(x^1), W(x^2), \cdots,$ *converges* a.s. *and* $W_\infty = \lim_{n\to\infty} W(x^n) = V(T(x))$ a.s. *for some unbiased* $x = (x_1, x_2, \cdots, x_k)$ *for which* $x_i \neq x_j$ *if* $i \neq j$.

THEOREM 2. *Let* $u_i^n = u_i(x^n)$ *and* $p_i^n = p(T_i(x^n))$; *then as* $m \to \infty$

$$m^{-1} \sum_{n=1}^{m} \sum_{i=1}^{k} p_i^n |x_i^n - u_i^n| \to 0 \qquad a.s. \tag{2.11}$$

MacQueen presents examples which show that the k means method *cannot* be counted on to provide minimum within class variance.

The k means method can be modified to increase or decrease the number of clusters under suitable conditions. Typically if a new point Z is too far from each of the reference points it can be made the first reference point of a $(k + 1)$st cluster (*refinement*). If two reference points are too close to each other one can combine their clusters by replacing the two reference points by a suitable weighted average (*coarsening*). The criteria for too far and too close can be set in advance by two parameters R for refinement and C for coarsening.

3. Distance considerations

The preceding section was based on the implicit assumption that Euclidean distance is the appropriate measure of distance. However it is known that in dealing with a random variable Z with a multivariate normal distribution with

mean μ and nonsingular covariance matrix A, the Mahalanobis distance measure

$$d(x, y) = (x - y)'A^{-1}(x - y) \tag{3.1}$$

is highly meaningful. The "ability" to test the hypothesis $H_1: \mu = \mu_1$ versus the alternative $H_2: \mu = \mu_2$ is an increasing function of $d(\mu_1, \mu_2)$. It is known that $d(Z, \mu)$ has the χ^2 (chi square) distribution with r degrees of freedom (d.f.) and that $d(Z, x)$ has the noncentral χ^2 distribution with r d.f. and noncentrality parameter $d(x, \mu)$.

It has been suggested that the Euclidean metric $(x - y)'(x - y)$ be replaced by the Mahalanobis metric in measuring the distance used in various cluster analysis techniques. A glance at Figure 1 indicates that this suggestion could lead to undesirable results.

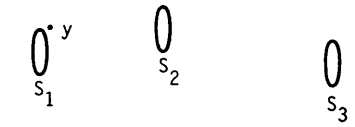

FIGURE 1

Figure 1 represents the diagram of three clusters in two dimensional space. Each cluster is described by a set of points lying in a thin ellipsoid and resembles a sample from a multivariate normal distribution. The three ellipsoids are similar with long vertical axes indicating that horizontal distances are important in determining the cluster to which a point belongs. On the other hand, the Mahalanobis distance using the covariance matrix for the *overall* set of points would tend to give most emphasis to the relatively unimportant vertical distances. The use of this metric might serve to persuade one to assign the point marked y to S_2 rather than to S_1.

It has been suggested that the above undesirable attribute be discounted by using Mahalanobis distance with a covariance matrix corresponding to within cluster variance rather than overall variance. That is, one should use

$$d^*(x, y) = (x - y)'A_w^{-1}(x - y), \tag{3.2}$$

where A_w is the within cluster covariance matrix

$$A_w = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \bar{x}_i)(x - \bar{x}_i)' \tag{3.3}$$

with the $S = (S_1, S_2, \cdots, S_k)$ representing a decomposition of the sample into k pairwise disjoint sets (clusters) whose averages are the $\bar{x}_i$.

Elaborations of this proposal have been treated by Friedman and Rubin [2] who consider the characteristic values and vectors of A_w with respect to the overall covariance matrix,

$$A_T = \sum_{x} (x - \bar{x})(x - \bar{x})'. \tag{3.4}$$

The potential disadvantage of using A_w becomes manifest in Figure 2 where two sharply defined clusters with almost singular covariance matrices combine to yield an A_w which is a multiple of the identity and corresponds to a multiple of the Euclidean metric. It is clear that for cluster 1 vertical distance is most important whereas for cluster 2 horizontal distance is crucial. To decide whether an arbitrary point belongs to S_1 or S_2, it seems most advisable to compare the appropriate metric in each case. Thus the point labeled y is more naturally associated with S_1 though it is closer to the center of S^2.

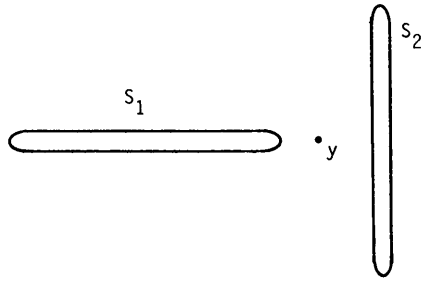

FIGURE 2

It is indicated in Figure 3 that two clusters could conceivably intersect with the result that each cluster effectively divides the other into two disconnected pieces. While this represents an undesirable, if uncommon situation the lack of connection hardly seems as serious a problem as the presence of the common part where points are difficult to classify into one cluster or the other. This more serious problem of difficulty in classification occurs often in less pathological appearing examples.

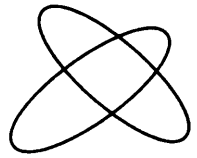

FIGURE 3

To overcome the difficulties rising from the above approaches we propose to introduce a modification of the k means method where each cluster determines its own metric, the Mahalanobis distance for that cluster. The possibility of using these metrics seems to have been previously considered only by Rohlf [4] and in connection with an hierarchical approach.

4. The modified procedure

Suppose that at the nth stage we have k clusters represented by reference points $x^n = (x_1^n, x_2^n, \cdots, x_k^n)$, weights $w^n = (w_1^n, w_2^n, \cdots, w_k^n)$ and covariances $A^n =$

$(A_1^n, A_2^n, \cdots, A_k^n)$. Corresponding to the reference points x and covariances A, we have a measurable decomposition of the r dimensional Euclidean space into $T(x, A) = \big(T_1(x, A), T_2(x, A), \cdots, T_k(x, A)\big)$, where $z \in T_i(x, A)$ only if

$$(z - x_i)'A_i^{-1}(z - x_i) \leqq (z - x_j)A_j^{-1}(z - x_j), \qquad 1 \leqq j \leqq k. \tag{4.1}$$

If a new observation Z is selected at random, we let $x_i^{n+1} = x_i^n$, $w_i^{n+1} = w_i^n$, and $A_i^{n+1} = A_i^n$ if $Z \notin T_i(x^n, A^n)$. If $Z \in T_i(x^n, A^n)$ we let

$$x_i^{n+1} = \frac{x_i^n w_i^n + Z}{w_i^n + 1}, \qquad w_i^{n+1} = w_i^n + 1 \tag{4.2}$$

and, setting $B_i^n = w_i^n A_i^n$, we let

$$B_i^{n+1} = B_i^n + \frac{w_i^n}{w_i^n + 1}(Z - x_i^n)(Z - x_i^n)', \tag{4.3}$$

$$B_i^{n+1} = B_i^n + \frac{w_i^n + 1}{w_i^n}(Z - x_i^{n+1})(Z - x_i^{n+1})'.$$

The motivation for this formula derives from the following algebra. Given m observations $y_1, y_2, \cdots, y_m$ with mean u_m let

$$B_m^* = mA_m^* = \sum_{\alpha=1}^{m}(y_\alpha - u_m)(y_\alpha - u_m)'. \tag{4.4}$$

Then

$$B_{m+1}^* = B_m^* + y_{m+1}y_{m+1}' + mu_m u_m' - (m + 1)u_{m+1}u_{m+1}', \tag{4.5}$$

where

$$(m + 1)u_{m+1} = mu_m + y_m. \tag{4.6}$$

It is easily seen that

$$\begin{aligned} B_{m+1}^* - B_m^* &= \frac{m}{m + 1}(y_{m+1} - u_m)(y_{m+1} - u_m)' \\ &= \frac{m + 1}{m}(y_{m+1} - u_{m+1})(y_{m+1} - u_{m+1})'. \end{aligned} \tag{4.7}$$

Thus, except for a scale factor, the covariance matrix A_m changes by the addition of a matrix of rank 1. This has a desirable aspect, for if C is nonsingular and

$$D = C + hh', \tag{4.8}$$

then

$$D^{-1} = C^{-1} - \frac{(C^{-1}h)(h'C^{-1})}{1 + h'C^{-1}h}. \tag{4.9}$$

Therefore the A_i^n can be inverted recursively.

Thus far we have an algorithm for applying the modified method once initial values x^1, w^1, A^1 are obtained. Some suggestion is necessary for initiating the iterative procedure. It is possible to take the first k observations as the components of x^1. To avoid singularity it is desirable to start with the A_i as positive definite symmetric matrices. These can be arbitrary, say the identity matrix. Alternatively some prior insights or previous information could lead to alternative suggestions. The value of w^1 could be $(1, 1, \cdots, 1)'$ or may also be assigned in a more or less arbitrary fashion. One can avoid unnecessarily rapid early fluctuations in A^n by starting w^1 with large components. Alternatively by selecting the B_i^1 to be relatively "large" in magnitude one can reduce the early fluctuations in A^n. An advantage of large B_i^1 with small w^1 over large w^1 is that the initial reference points are not given undue weight.

5. Comments

5.1 *Economic variations.* The computational cost per iteration of the proposed modification is greater than that for the Euclidean metric version by an amount necessary to compute the revised A_i^{-1} and the k distances. This extra cost is of the order of magnitude of kr^2 where r is the dimensionality of the space. If r is large the extra cost may necessitate the use of some short cuts.

One possibility that may be worth exploring is to decompose A_i into principal components. Then one can confine attention to a smaller dimensional space which is spanned by characteristic vectors corresponding to a few of the largest characteristic values from each of the A_i. Thus using three vectors from each of five groups yields a 15 dimensional space which leads to a considerable savings if r is say 50. To use this technique effectively one would have to carry out a substantial number of iterations with a given subspace before recomputing the principal components.

A variation of this approach is to separate the characteristic roots of A_i into two groups and to approximate distances. To illustrate suppose A_1 has characteristic roots $\lambda_1 \geqq \lambda_2 \geqq \cdots \lambda_{r_1} \geqq \lambda_{r_1+1} \geqq \cdots \geqq \lambda_r$ and corresponding vectors $u_1, u_2, \cdots, u_r$. Then an arbitrary vector z may be decomposed so that $z = \Sigma z_i u_i$ where the z_i are the projections of z on u_i and the distance $z'A_1^{-1}z = \Sigma \lambda_i^{-1} z_i^2$. This may be approximated (from below) by

$$\sum_{i=1}^{r} \lambda_i^{-1} z_i^2 + \lambda_{r_1+1}^{-1} \sum_{i=r_1+1}^{r} z_i^2 = \sum_{i=1}^{r_1} \lambda_i^{-1} z_i^2 + \lambda_{r_1+1}^{-1} \left(\sum_{i=1}^{r} z_i^2 - \sum_{i=1}^{r_1} z_i^2 \right). \tag{5.1}$$

Thus all that are required are the lengths of the projections of z on the first r_1 characteristic vectors and the Euclidean length of z. With this variation it is also desirable to use the λ_i and u_i for several iterations before recomputing.

One could elaborate on this variation by dividing the λ_i into several groups. It is questionable that this would help much.

It should be remarked that as the dimensionality of the space increases, the sample size necessary to obtain reliable results tends to increase. At this point it would be difficult to anticipate the extent.

5.2. *Singular covariances.* If some of the clusters lie in proper linear subspaces of the Euclidean space, particularly if they lie in different subspaces, the ability to distinguish the clusters should be very great. In practice this desirable situation means that we deal with covariance matrices which are singular or almost so. Initiating the iteration with nonsingular matrices A_i leads, through equation (4.3) to a sequence of matrices which are nonsingular. However, in the above described situation, these matrices should tend to be successively closer to singular matrices. While no enormous difficulty is anticipated in the inversion through equation (4.9) one should be prepared to recognize the phenomenon as it develops, since its presence points to potentially important properties of the data.

5.3. *Relevance of modified approach.* The situation in which the procedures described seem most relevant is when the population consists of a set of clusters each of which resembles the ellipsoidal form of a multivariate normal distribution and the covariance structures of these clusters are substantially different and preferably confined to different linear subspaces. Clusters which tend to curve, particularly those cases where some points of one cluster tend to be closer to points of another cluster than to other points of the same cluster (see Figure 4), should not yield much information to our modified approach.

FIGURE 4

5.4. *Coarsening and refining.* Assuming multivariate normal distributions one would expect that $(Z - x_i^n)'A_i^n(Z - x_i^n)$ would have the chi square distribution if z is in the ith cluster. Thus large or small values of this statistic could be used for coarsening and refining the clusters where the discrimination between large and small should relate in part with the percentiles of the chi square distribution. Such a procedure is recommended here with reservations since nonnormal behavior will alter the distribution of $(Z - x_i^n)'A_i(Z - x_i^n)$ and it seems to be more conservative to keep track of the empirical distribution of means and variances of substantial numbers of recent values of these distances to see what is unusually large or small.

5.5. *What is a cluster?* The k means method is an approach to cluster analysis based mainly on the "metric" concept of a cluster as a set of points which are closer to one another than to other points. Another general approach which requires more calculation, if one does not use sampling creatively, is to regard a cluster as a set of points each of which has a nearby neighbor of the cluster.

Although in most illustrative examples in two dimensional space the informal application of both approaches yields the same common sense idea of cluster; this is not necessarily the case. It is important to prove some theorems which will establish properties of this distribution p which will imply that both approaches yield the same results. Otherwise one must wonder whether one computationally convenient method will yield unrecognizable clusters from another point of view. Part of the weakness of the highly nontrivial MacQueen conclusions derives in part from requiring too general a domain of applicability. It may be easier to get stronger results for those distributions p for which various approaches coincide.

REFERENCES

[1] E. FORGY, "Cluster analysis of multivariate data: Efficiency versus interpretability of classification abstract," *Biometrics*, Vol. 21 (1965), p. 768.

[2] H. R. FRIEDMAN and J. RUBIN, "On some invariant criteria for grouping data," *J. Amer. Statist. Assoc.*, Vol. 62 (1967), pp. 1159–1178.

[3] J. MACQUEEN, "Some methods for classification and analysis on multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1966, Vol. 1, pp. 281–297.

[4] F. J. ROHLF, "Adaptive hierarchical clustering schemes," *Syst. Zool.*, Vol. 19 (1970), pp. 58–83.