IBM Technical Disclosure Bulletin Vol. 18 No. 8 January 1976

CLUSTER ANALYSIS

R. Capra, A. Lena, U. Santarelli and P. Vescovi

Cluster Analysis of a multivariate sample X:

$$X = \{\underline{x}_i \mid \underline{x}_i \in R^L,\ i=1,..,N\}$$

is the partitioning of X into M homogeneous groups according to a minimization criterion E based on the distance d:

$$d : R^L \times R^L \to R^+$$

The algorithm elaborated has the following operational characteristics:

- Sequential organization of data.
- No interdistance matrix.
- Processing times linearly dependent on N.
- No effective limitation on N, L, M.
- Euclidean distance.

Moreover, in the Fortran implementation of the algorithm, it is possible to have:

- Standardization of the input data (in order to avoid scale effect).
- Weighting of the variables (in order to force a scale effect).

The minimization criterion used to measure the nonhomogeneity of a partition of X into groups $X_1\ X_2\ \ldots\ X_M$ is given by the function:

$$E = \sum_{k}^{M}{}_{1} E_k$$

where;

$$E_k = \sum_{\underline{x}_i \in X_k} d^2\ (\underline{x}_i, \underline{b}_k)$$

and,

$$\underline{b}_k = \frac{1}{n_k} \sum_{\underline{x}_i \in X_k} \underline{x}_i$$

This function, in addition to its clear geometric significance - total sum of the moments of inertia of the systems of points in relation to its own barycenter $\underline{b}_k$ - has useful computational properties.

It can be seen that if $\underline{x}_i \in X_k$, we have:

$$E_k^- = E_k(X_k) - E_k'(X_k - \{\underline{x}_i\}) = \frac{n_k}{n_k - 1} d^2(\underline{x}_i, \underline{b}_k)$$

Likewise, if $\underline{x}_i \notin X_k$, we have:

$$E_k^+ = E_k'(X_k \cup \{\underline{x}_i\}) - E_k(X_k) = \frac{n_k}{n_k + 1} d^2(\underline{x}_i, \underline{b}_k)$$

It follows that, for each partition, the ΔE variation of E, brought about by the shifting of one element $\underline{x}_i$ from group X_k to group X_h:

$$\Delta E = E_h^+ - E_k^-$$

depends only on:

$$\underline{x}_i, \underline{b}_k, n_k, \underline{b}_h, n_h.$$

In this way it is possible to have in core only one observation at once.

A partition is considered stable when, for every $\underline{x}_i$ of X, we have:

$$\Delta E > 0$$

In this case, we see that if $\underline{x}_i \in X_k$:

$$d(\underline{x}_i, \underline{b}_k) \leq \min d(\underline{x}_i, \underline{b}_h') \quad \text{for } h \neq k,$$

with $\underline{b}_h'$ barycenter of $X_h \cup \{\underline{x}_i\}$.

Since is not possible to guarantee that a stable partition is the best, because of the dependence of the classification on the order in

which the data are presented, the algorithm envisages the subsequent application of a C procedure so as to reach a solution independent of the initial sequence of the data while maintaining performance unchanged in relation to the E criterion.

The C procedure sorts the set $Y = \{\underline{Y}_i \mid \underline{Y}_i \in R^L, i=1,..,N\}$

into ng groups, utilizing the initial ng barycenters of B_o:

$$B_o = \{\underline{b}_1^o, \ldots, \underline{b}_{ng}^o\},$$

as specified by an Isw switch, and placing in B,

$$B = \{\underline{b}_1, \ldots \underline{b}_{ng}\},$$

the barycenters of the classification is obtained.

This gives:

$$B \leftarrow C\ (Y, ng, B_o, Isw).$$

C operates in the following way:

a) Make $B=B_o$ and $n_k=0, Y_k=\Phi, k=1,\ldots,ng$

b) If Isw=0 go to e)

c) Compute $d(\underline{Y}_i, \underline{b}_k^o) = \min_{h \neq k} d(\underline{Y}_i, \underline{b}_h^o)$ for every i

make $Y_k = Y_k \cup \{\underline{Y}_i\}$; make $n_k = n_k + 1$

d) Compute $\underline{b}_k = \frac{1}{n_k} \Sigma \underline{Y}_i$ for every k

e) Compute

$$E_k^- = \frac{n_k}{n_k - 1}\ d^2\ (\underline{Y}_i, \underline{b}_k)$$ for every i

$$E_h^+ = \frac{n_h}{n_h + 1}\ d^2\ (\underline{Y}_i, \underline{b}_h)$$ for every i and for $h \neq k$

If $E_k^- \leq \min_h E_h^+$, $\underline{Y}_i$ remains in Y_k

If $E_{h_o}^+ = \min_h E_h^+$ and $E_{h_o}^+ < E_k^-$ move $\underline{Y}_i$

from Y_k to Y_{h_o} and make:

$Y_k = Y_k - \{\underline{y}_1\}$; $\underline{b}_k = (\underline{b}_k n_k - \underline{y}_1)/(n_k - 1)$; $n_k = n_k - 1$;

$Y_{h_o} = y_{h_o} \cup \{\underline{y}_1\}$; $\underline{b}_{h_o} = (\underline{b}_{h_o} n_{h_o} + \underline{y}_1)/(n_{h_o} + 1)$; $n_{h_o} = n_{h_o} + 1$;

f) If any $\underline{y}_1$ has been moved (E is reduced) go to e); otherwise stop.

The algorithm works in the following way:

a) Make $ng = \sqrt{MN}$, $B_o = \{\underline{0}, \underline{0}, \ldots \underline{0}\}$

b) Compute $B' \leftarrow C(X', ng, B_o, 0)$

where X' is possibly a subsample of X if N/M is considered excessive depending on economic considerations.

c) Make ng=M.

Choose $\bar{B} \subset B'$ card $(\bar{B}) = ng$ in the following way:

1) $k = 1$; $\bar{B} = (\underline{b}'_1)$ 2) $k = 2,;,ng$; compute:

$$d(\underline{b}'_{i_o}, \bar{\underline{b}}_{j_o}) = \max_{\underline{b}'_i \in B' - \bar{B}} \min_{\bar{\underline{b}}_j \in \bar{B}} d(\underline{b}'_i, \bar{\underline{b}}_j); \quad \bar{B} = \bar{B} \cup (\underline{b}'_{i_o}).$$

This makes it possible to choose the sparsest subset of M barycenters in B'.

d) $B'' \leftarrow C(B', M, \bar{B}, 0)$

e) $B \leftarrow C(X, M, B'', 1)$

which performs the final classification on the entire X sample, utilizing B'' as seeds of the classification.

Because of the absence of specific tests to evaluate the performances of Cluster Analysis Algorithms, tests were based upon the comparison of the automatic solution and the subjective classification of known samples.

The similarity between the clustering results achieved by the two methods confirms the reliability of the algorithm, at least when it is possible to give a coherent Euclidean interpretation of the subjective similarity criterion.