K–means clustering in a low-dimensional Euclidean space

Geert De Soete[1] & J. Douglas Carroll[2]

[1] Department of Data Analysis, University of Ghent,
Henri Dunantlaan 2, B-9000 Ghent, Belgium

[2] Graduate School of Management, Rutgers University,
81 New Street, Newark, New Jersey 07102–1895, U.S.A.

Summary: A procedure is developed for clustering objects in a low-dimensional subspace of the column space of an objects by variables data matrix. The method is based on the K–means criterion and seeks the subspace that is maximally informative about the clustering structure in the data. In this low-dimensional representation, the objects, the variables and the cluster centroids are displayed jointly. The advantages of the new method are discussed, an efficient alternating least-squares algorithm is described, and the procedure is illustrated on some artificial data.

1. Introduction

Let $\mathbf{X} = ((x_{ij}))$ contain measurements of N objects on M variables. It will be assumed that the columns of $\mathbf{X}$ are centered and suitably normalized. A common purpose in data analysis entails clustering the N objects into K ($K \ll N$) homogeneous clusters, starting from $\mathbf{X}$. When it is suspected that some of the variables do not contribute much to the clustering structure in the data, or when the number of variables is large, researchers often carry out a preliminary principal component or factor analysis on $\mathbf{X}$. The cluster analysis is then performed on the object scores on the first few components or factors. This two-step procedure, called "tandem clustering" by Arabie and Hubert (in press), has been warned against by several authors (e.g., Arabie and Hubert (in press), Chang (1983), DeSarbo et al. (1990)), because the first few principal components or factors of $\mathbf{X}$ do not necessarily define a subspace that is most informative about the cluster structure in the data. Nevertheless, tandem clustering is still widely used in some disciplines, such as marketing research (e.g., Doyle and Saunders (1985), Furse et al. (1984)).

However, including variables that do not contribute much to the clustering structure in the data might obscure the clustering structure or mask it altogether (e.g., Milligan (1980)). Also, a low-dimensional (say two- or three-dimensional) representation of the cluster structure in $\mathbf{X}$ is very useful. Such a representation can aid the researcher in evaluating and interpreting the results of the cluster analysis. For instance, it would allow the researcher to visually inspect the shape of the clusters, to identify outliers, and to assess the degree to which each variable contributes to the cluster structure in the data. Hence, what is needed is a procedure that constructs a low-dimensional representation of the data such that the clustering structure in the data is maximally revealed. However, instead of a two-step tandem clustering procedure, a method is called for that simultaneously performs the cluster analysis and the dimension reduction. In this paper, such a method is developed based on the K–means criterion (MacQueen, 1967). While simultaneous clustering and dimension reduction methods have been recently developed in the context of multidimensional scaling (DeSarbo et al. (1990, 1991), De Soete and Heiser (1993), De Soete and Winsberg (1993), Heiser

E. Diday et al. (eds.), *New Approaches in Classification and Data Analysis*

(1993), Winsberg and De Soete (1993)), procedures that can be applied to a general objects × variables data matrix are rare (but see, Van Buuren and Heiser (1989)). In this paper, a new approach to devising such a procedure is proposed. In Section 2, we describe the model on which the new procedure is based. Next, in Section 3, an efficient algorithm is developed for fitting the model. An illustrative application is presented in Section 4 and some concluding remarks are offered in the final section.

2. Model

Let the $K \times M$ matrix $\mathbf{C}$ contain the centroids of the K clusters and let $\mathbf{E}$ be an $N \times K$ binary indicator matrix that specifies for each object cluster membership:

$$e_{ik} = \begin{cases} 1 & \textit{iff} \text{ object } i \text{ belongs to cluster } k, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Since in K–means cluster analysis each object is assigned to one and only one cluster,

$$\sum_{k=1}^{K} e_{ik} = 1 \tag{2}$$

holds for $i = 1, \ldots, N$.

In K–means clustering the matrices $\mathbf{E}$ and $\mathbf{C}$ are determined such that the sum of squared Euclidean distances between the objects and the centroids of the clusters to which they belong, is minimal. That is, $\mathbf{E}$ and $\mathbf{C}$ are sought such that the least-squares loss function

$$\begin{aligned} F(\mathbf{E}, \mathbf{C}) &= ||\mathbf{X} - \mathbf{EC}||^2 \\ &= \text{tr}(\mathbf{X} - \mathbf{EC})'(\mathbf{X} - \mathbf{EC}) \end{aligned} \tag{3}$$

is minimal.

When a K–means clustering in a low-dimensional space is desired, the K centroids are required to lie in an R–dimensional subspace of the column space of $\mathbf{X}$. When the data are column-centered, the K cluster centroids always define a $(K-1)$–dimensional subspace. Thus, to achieve dimension reduction, R should be smaller than $\min(K-1, M)$. The K centroids can be restricted to lie in an R–dimensional subspace by imposing the constraint

$$\text{rank}(\mathbf{C}) = R \tag{4}$$

on $\mathbf{C}$.

The K–means clustering procedure in an R–dimensional space then amounts to determining $\mathbf{E}$ and $\mathbf{C}$ such that (3) is minimized subject to (2), (4) and $e_{ik} \in \{0, 1\}$. In the next section, an algorithm is presented for solving this constrained optimization problem.

3. Method

3.1 Algorithm

The loss function (3) can be written as

$$F(\mathbf{E}, \mathbf{C}) = \sum_{i=1}^{N} \sum_{j=1}^{M} \left(x_{ij} - \sum_{k=1}^{K} e_{ik} c_{kj} \right)^2$$