k-Means Clustering with Outlier Detection, Mixed Variables and Missing Values

D. Wishart

Department of Management,
University of St. Andrews, St. Katharine's West,
The Scores, St. Andrews, Fife KY16 9AL, Scotland
Email: d.wishart@st-andrews.ac.uk
Website: www.clustan.com
Tel: +44 131 337 1448

Abstract: This paper addresses practical issues in k-means cluster analysis or segmentation with mixed types of variables and missing values. A more general k-means clustering procedure is developed that is suitable for use with very large datasets, such as arise in data mining and survey analysis. An exact assignment test guarantees that the algorithm will converge, and the detection of outliers allows the densest regions of the sample space to be mapped by tessellations of tightly-specified spherical clusters. A summary tree is obtained for the resulting k-cluster partition.

1 Introduction

Clustering problems routinely occur in survey analysis and data mining where incomplete data and different types of variables are present. A popular method of classification is k-means analysis, which partitions a set of cases into k clusters so as to minimise the "error" or sum of squared distances of the cases about the cluster means. However, k-means analysis is usually only implemented with quantitative variables and complete data. This paper extends Gower's General Similarity Coefficient to k-means analysis so that mixed data types, missing values and differential weighting of cases or variables can be handled, with an exact assignment test that guarantees the algorithm will converge.

2 k-Means Cluster Analysis

The conventional k-means clustering procedure is as follows:

1. Choose an initial partition of the cases into k clusters. This can be a tree section, k selected cases, or a random assignment to k clusters.
2. Compute the distance from every case to the mean of each cluster, and assign the cases to their nearest clusters.
3. Re-compute the cluster means following any change of cluster membership at step 2.

M. Schwaiger et al. (eds.), *Exploratory Data Analysis in Empirical Research*

4. Repeat steps 2 and 3 until no further changes of cluster membership occur in a complete iteration. The procedure has now converged to a stable k-cluster partition.

The k-means procedure was first proposed by Thorndike (1953) in terms of minimizing the average of all the within-cluster distances. Forgey (1965), Ball (1965), Jancey (1966), Ball and Hall (1967), MacQueen (1967) and Beale (1969) implemented it computationally to minimise the distances from the cluster means. It is provided with Clustan (Wishart (1970, 1984, 1999)).

Some programs re-compute the cluster means only at the end of each completed iteration, after all changes in cluster membership have been made, and not following each change of cluster membership as in step 3. This has the advantage that the resulting partition is independent of the case order; but it is generally slower to converge than the progressive re-calculation of means at step 3, sometimes referred to as "drifting means".

The object of k-means analysis is to arrive at a stable k-cluster partition in which each case is closest to the mean of the cluster to which it is assigned. In essence, the "model" is the final set of k cluster means, and the procedure seeks to minimise the "error" in the model, as measured by the sum of the squared distances from the cases to the cluster means. Some authors advocate imputing missing values prior to clustering, because their programs only work with complete data. However, the k-means algorithm developed below allows for missing values by estimating cluster means, distances to cluster means, and criterion function values from the complete data, ignoring any missing values.

3 General Similarity Coefficient

The starting point is a General Similarity Coefficient s_{ij} for the comparison of two cases i and j, proposed by Gower (1971) that has been widely implemented and used:

$$s_{ij} = \sum_k w_{ijk} s_{ijk} \Big/ \sum_k w_{ijk} \tag{1}$$

where s_{ijk} is a similarity component for the k^{th} variable (defined below), and the weight w_{ijk} is 1 if the comparison is valid for the k^{th} variable, or 0 otherwise. Thus $w_{ijk} = 0$ where one or both of the observations on the k^{th} variable is missing.

For quantitative variables, Gower standardises s_{ijk} by the range r_k of the k^{th} variable: