An optimising approach to alternative clustering schemes

Alan Jessop

Abstract Clustering objects into groups is usually done using a statistical heuristic or an optimisation. The method depends on the size of the problem and its purpose. There may exist a number of partitions which do not differ significantly but some of which may be preferable (or equally good) when aspects of the problem not formally contained in the model are considered in the interpretation of the result. To decide between a number of good partitions they must first be enumerated and this may be done by using a number of different heuristics. In this paper an alternative method is described which uses an integer linear programming model having the number and size distribution of groups as objectives and the criteria for group membership as constraints. The model is applied to three problems each having a different measure of dissimilarity between objects and so different membership criteria. In each case a number of optimal solutions are found and expressed in two parts: a core of groups, the membership of which does not change, and the remaining objects which augment the core. The core is found to contain over three quarters of the objects and so provides a stable base for cluster definition.

Keywords ILP · Multicriteria · Statistics · Cluster

JEL Classification C61

1 Introduction

There are many circumstances in which objects are partitioned into groups: task partitioning, the analysis of social relations, the definition of taxa, and others.
Objects which are similar or close will be in the same group. The necessary pairwise relations may be given directly, as when pairs of people in a society are defined as being linked in some way, or each object may be described according to a number of attributes and from these data a measure of the dissimilarity between each pair is calculated. In these latter problems the groups are usually called clusters. For the development of our model we consider that a grouping scheme will have four constituents:

- \( d \): a measure of pairwise dissimilarity
- \( c \): a criterion for group membership (based on \( d \))
- \( n \): the number of groups
- \( u \): the unevenness of the size distribution of groups

There are a number of measures of dissimilarity (e.g. Everitt 1993). Some measure a distance between a pair of objects based on a number of attributes possessed by each; Euclidean distance, for instance. Other measures describe the distance between pairs of groups, the mean of the pairwise Euclidean distances. The purpose is to have some metric such that the larger the value the less similar are the pair and so the less the justification for including them in the same group. This allows object pairs to be defined as either too dissimilar to be grouped or sufficiently similar to form candidate groups from which a partition may be formed.

It is common that in cluster formation a hierarchical heuristic is used in which, for instance, objects join the nearest cluster until all have been allocated. The process of cluster formation is displayed as a tree showing the level of dissimilarity at which clusters are formed. In the light of contextual or theoretical considerations this tree is inspected and a decision made as to the most appropriate partition. Once \( d \) has been chosen \( c \), \( n \) and \( u \) are considered together in interpreting the result. This may be cognitively difficult. The result is a single partition. Some form of sensitivity testing, by choosing different measures for \( d \) say, or by using more than one method to form clusters (Kaufman and Rousseeuw 1990) is recommended as a way of exploring alternatives. Cluster formation is, in this broad sense, interactive.

Task variety may be reduced by pre-setting the number of groups to \( n = k \) (the \( k \)-means method). There may be no strong reason for preferring a particular \( k \) so that while this method reduces the number of factors to be considered in interpreting results it requires a corresponding increase in sensitivity testing to see the effect of different values of \( k \). Nonetheless, structuring the analysis in this way offers the prospect of easier interpretation of results.

Optimization has been intermittently proposed for clustering problems using aggregate dissimilarity as an objective. For example, we may minimise the sum of squares of intra-cluster dissimilarities given appropriate constraints on the number of groups or their size. Clusters are formed by the use of mathematical programming methods (Rao 1971; Hansen and Jaumard 1997), notably linear programming (Vinod 1969; Joseph and Bryson 1998) and dynamic programming (Jensen 1969; van Os and Meulman 2004). A considerable benefit of optimising approaches is that they have a clear criterion which allows for an assessment of how good optimal and other partitions are (Li 2006). This helps the evaluation of alternative partitions. These optimising methods work well for small and medium sized problems but they become infeasible