Cluster Analysis Using Rough Clustering and k–Means Clustering

Kevin E. Voges
University of Canterbury, New Zealand

INTRODUCTION

Cluster analysis is a fundamental data reduction technique used in the physical and social sciences. The technique is of interest to managers in information science because of its potential use in identifying user needs though segmenting users such as Web site visitors. In addition, the theory of rough sets is the subject of intense interest in computational intelligence research. The extension of this theory into rough clustering provides an important and potentially useful addition to the range of cluster analysis techniques available to the manager.

Cluster analysis is defined as the grouping of "individuals or objects into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters" (Hair, Anderson, Tatham & Black, 1998, p. 470). There are a number of comprehensive introductions to cluster analysis (Arabie, Hubert & De Soete, 1994; Cramer, 2003; Everitt, Landau & Leese, 2001). Techniques are often classified as hierarchical or nonhierarchical (Hair et al., 1998), and the most commonly used nonhierarchical technique is the *k*-means approach developed by MacQueen (1967). Recently, techniques based on developments in computational intelligence have also been used as clustering algorithms. For example, the theory of fuzzy sets developed by Zadeh (1965), which introduced the concept of partial set membership, has been applied to clustering (Dumitrescu, Lazzerini & Jain, 2000). Another technique receiving considerable attention is the theory of rough sets (Pawlak, 1982), which has led to clustering algorithms referred to as rough clustering (do Prado, Engel & Filho, 2002; Voges, Pope & Brown, 2002).

This article provides brief introductions to *k*-means cluster analysis, rough sets theory, and rough clustering, and compares *k*-means clustering and rough clustering. The article shows that rough clustering provides a more flexible solution to the clustering problem, and can be conceptualized as extracting *concepts* from the data, rather than strictly delineated subgroupings (Pawlak, 1991). Traditional clustering methods generate *extensional* descriptions of groups (i.e., which objects are members of each cluster), whereas clustering techniques based on rough sets theory generate *intensional* descriptions (i.e., what are the main characteristics of each cluster) (do Prado et al., 2002). These different goals suggest that both *k*-means clustering and rough clustering have their place in the data analyst's and the information manager's toolbox.

BACKGROUND

k-Means Cluster Analysis

In the *k*-means approach, the number of clusters (*k*) in each partition of the data set is decided *prior to* the analysis, and data points are randomly selected as the initial estimates of the cluster centres (referred to as centroids). The remaining data points are assigned to the closest centroid on the basis of the distance between them, usually using a Euclidean distance measure. The aim is to obtain maximal homogeneity within clusters (i.e., members of the same cluster are most similar to each other), and maximal heterogeneity between clusters (i.e., members of different clusters are most dissimilar to each other).

K-means cluster analysis has been shown to be quite robust (Punj & Stewart, 1983). Despite this, the approach suffers from many of the problems associated with all traditional multivariate statistical analysis methods. These methods were developed for use with variables that are normally distributed and that have an equal variance-covariance matrix in all groups. In most realistic data sets, neither of these conditions necessarily holds.

Rough Sets

The concept of rough sets (also known as approximation sets) was introduced by Pawlak (1982, 1991), and is based on the assumption that with every record in the information system (the data matrix in traditional data analysis terms), there is associated a certain amount of information. This information is expressed by means of attributes (variables in traditional data analysis terms), used as descriptions of the objects. For example, objects could be individual users in a study of user needs, and attributes could be characteristics of the users such as gender, level of experience, age, or other characteristics considered relevant. See Pawlak (1991) or Munakata (1998) for comprehensive introductions.

In rough set theory, the data matrix is represented as a table, the information system. The complete information system expresses all the knowledge available about the objects being studied. More formally, the information system is a pair, $S = (U, A)$, where U is a non-empty finite set of objects called the universe and $A = \{ a1, ..., aj \}$ is a non-empty finite set of attributes describing the objects in U. With every attribute $a \in A$ we associate a set Va such that $a : U \rightarrow Va$. The set Va is called the domain or value set of a. In traditional data analysis terms, these are the values that each variable can take (e.g., gender can be male or female; users can have varying levels of experience).

A core concept of rough sets is that of indiscernibility. Two objects in the information system about which we have the same knowledge are indiscernible. Let $S = (U, A)$ be an information system; then with any subset of attributes B, ($B \subseteq A$), there is associated an equivalence relation, $IND_A(B)$, called the B-indiscernibility relation. It is defined as:

$$IND_A(B) = \{ (x, x') \in U^2 \mid \forall a \in B \; a(x) = a(x') \}$$

In other words, for any two objects (x and x') being considered from the complete data set, if any attribute a, from the subset of attributes B, is the same for both objects, they are indiscernible (on that attribute). If (x, x') $\in IND_A(B)$, then the objects x and x' are indiscernible from each other when considering the subset B of attributes.

Equivalence relations lead to the universe being divided into partitions, which can then be used to build new subsets of the universe. Two of these subsets of particular use in rough sets theory are the lower approximation and the upper approximation. Let $S = (U, A)$ be an information system, and let $B \subseteq A$ and $X \subseteq U$. We can describe the set X using only the information contained in the attribute values from B by constructing the B-lower and B-upper approximations of X, denoted $B_*(X)$ and $B^*(X)$ respectively, where:

$$B_*(X) = \{ x \mid [x]_B \subseteq X \}, \text{ and } B^*(X) = \{ x \mid [x]_B \cap X \neq \varnothing \}$$

The set $BN_B(X)$ is referred to as the boundary region of X, and is defined as the difference between the upper approximation and the lower approximation. That is:

$$BN_B(X) = B^*(X) - B_*(X)$$

If the boundary region of X is the empty set, then X is a crisp (exact) set with respect to B. If the boundary region is not empty, X is referred to as a rough (inexact) set with respect to B. The important insight of Pawlak's work is his definition of a set in terms of these two sets, the lower approximation and the upper approximation. This extends the standard definition of a set in a fundamentally important way.

Rough Clustering

Rough clusters are a simple extension of the notion of rough sets. The value set (Va) is ordered, which allows a measure of the distance between each object to be defined, and clusters of objects are then formed on the basis of their distance from each other. An object can belong to more than one cluster. Clusters can then be defined by a lower approximation (objects exclusive to that cluster) and an upper approximation (all objects in the cluster which are also members of other clusters), in a similar manner to rough sets.

Let $S = (U, A)$ be an information system, where U is a non-empty finite set of M objects $(1 \leq i \leq M)$, and A is a non-empty finite set of N attributes $(1 \leq j \leq N)$ on U. The j^{th} attribute of the i^{th} object has value $R(i, j)$ drawn from the ordered value set Va.

For any pair of objects, p and q, the distance between the objects is defined as:

$$D(p,q) = \sum_{j=1}^{N} | R(p,j) - R(q,j) |$$

That is, the absolute differences between the values for each object pair's attributes are summed. The distance measure ranges from 0 (indicating indiscernible objects) to a maximum determined by the number of attributes and the size of the value set for each attribute.

One algorithm for producing rough clusters is as follows (Voges et al., 2002). Initially, a distance matrix for all paired object comparisons is calculated. All object pairs at interobject distance D, where D steps from 0 to a determined maximum, are identified. Each object pair (a_i, a_j) can be in one of three situations in relation to current cluster membership, with the following consequences:

1. Both objects have not been assigned to any prior cluster. A new cluster is started with a_i and a_j as the first members.
2. Both objects are currently assigned to clusters. Object a_i is assigned to object a_j's earliest cluster, and object a_j is assigned to object a_i's earliest cluster. The earliest cluster is the first cluster the object was assigned to.
3. One object, a_i, is assigned to a cluster and the other object, a_j, is not assigned a cluster. Object a_j is assigned to object a_i's earliest cluster.