Cluster Analysis Using Rough Clustering and K-Means Clustering

D

Kevin E. Voges
University of Canterbury, New Zealand

INTRODUCTION

Cluster analysis is a fundamental data reduction technique used in both the physical and social sciences. The extension of Rough Sets theory into cluster analysis through the techniques of Rough Clustering provides an important and potentially useful addition to the range of cluster analysis techniques available to the manager and the researcher.

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, Black, Babin, & Anderson, 2009). There are a number of comprehensive introductions to cluster analysis (Abonyi & Feil, 2007; Arabie, Hubert & De Soete, 1996; Cramer, 2003; Everitt, Landau, Leese, & Stahl, 2011; Gan, Ma, & Wu, 2007). Techniques are often classified as hierarchical or nonhierarchical (Hair et al., 2009), and the most commonly used nonhierarchical technique is the *k*-means approach developed by MacQueen (1967). Over the past few decades, techniques based on developments in computational intelligence have been used as clustering algorithms. For example, the theory of fuzzy sets developed by Zadeh (1965), who introduced the concept of partial set membership, has been applied to clustering (Abonyi & Feil, 2007; Dumitrescu, Lazzerini, & Jain, 2000).

Fuzzy clustering has developed an extensive literature, too broad to be thoroughly reviewed here. However, two extensions will be briefly considered to demonstrate the flexibility of the technique. Atanassov (1986) extended Zadeh's fuzzy set to a general form called an intuitionistic fuzzy set (IFS), which has been found to be more useful in dealing with uncertainty than a standard fuzzy set. Xu, Chen and Wu (2008) report an application of this IFS concept to clustering. In a second extension, Dunn (1973), and Bezdek (1981) proposed a Fuzzy C-means technique (FCM), which is one of the most commonly used objective function-based clustering techniques. Instead of assigning each object to a single cluster, class membership is relaxed by computing the membership grades using a unit interval. As will be seen below, this has similarities to clustering using rough sets. Izakian and Pedrycz (2014) developed an extension to the FCM, where the distance function is given adjustable weight parameters, quantifying the impact coming from blocks of features rather than from individual features. They also show the increased use of hybridization techniques (explored later in this article), using particle swarm optimization to optimize the weights. Genetic algorithms have also been applied to clustering tasks (Maulik, Bandyopadhyay, & Mukhopadhyay, 2011).

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; Kumar, Krishna, Bapi, & De, 2007; Lingras & Peters, 2011; Parmar, Wu, & Blackhurst, 2007; 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.

DOI: 10.4018/978-1-4666-5888-2.ch160

Copyright © 2015, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

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), but has many limitations. The method was developed for use with normally distributed variables 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, 2002), 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 $A = \{ a1, \ldots, aj \}$ 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, INDA (B), called the B-indiscernibility relation. It is defined as:

$$\text{IND}A\ (B) = \{\ (x, x') \in U^2 \mid \forall\ a \in B\ a(x) = a(x')\ \}$$

That is, 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') $\in$ INDA (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. Pawlak's insight was to define a set in terms of these two sets, the lower approximation and the upper approximation, which extended the standard definition of a set in a fundamentally important way.

5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/cluster-analysis-using-rough-clustering-and-k-means-clustering/112572

Related Content

Information Technology / Systems Offshore Outsourcing: Key Risks and Success Factors
Mahesh S. Raisinghani, Brandi Starr, Blake Hickerson, Marshelle Morrisonand Michael Howard (2010). *Breakthrough Discoveries in Information Technology Research: Advancing Trends (pp. 1-21).*
www.irma-international.org/chapter/information-technology-systems-offshore-outsourcing/39567

Fuzzy Decision Support System for Coronary Artery Disease Diagnosis Based on Rough Set Theory
Noor Akhmad Setiawan (2014). *International Journal of Rough Sets and Data Analysis (pp. 65-80).*
www.irma-international.org/article/fuzzy-decision-support-system-for-coronary-artery-disease-diagnosis-based-on-rough-set-theory/111313

Interview: The Systems View from Barry G. Silverman: A Systems Scientist
Manuel Moraand Miroljub Kljajic (2010). *International Journal of Information Technologies and Systems Approach (pp. 57-63).*
www.irma-international.org/article/interview-systems-view-barry-silverman/45161

Fuzzy Decision Support System for Coronary Artery Disease Diagnosis Based on Rough Set Theory
Noor Akhmad Setiawan (2014). *International Journal of Rough Sets and Data Analysis (pp. 65-80).*
www.irma-international.org/article/fuzzy-decision-support-system-for-coronary-artery-disease-diagnosis-based-on-rough-set-theory/111313

Serious Games and the Technology of Engaging Information
Peter A. Smithand Clint Bowers (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 2591-2599).*
www.irma-international.org/chapter/serious-games-and-the-technology-of-engaging-information/112675