Interval Set Representations of Clusters

Pawan Lingras
Saint Mary's University, Canada

Rui Yan
Saint Mary's University, Canada

Mofreh Hogo
Czech Technical University, Czech Republic

Chad West
IBM Canada Limited, Canada

INTRODUCTION

The amount of information that is available in the new information age has made it necessary to consider various summarization techniques. Classification, clustering, and association are three important data-mining features. Association is concerned with finding the likelihood of co-occurrence of two different concepts. For example, the likelihood of a banana purchase given that a shopper has bought a cake. Classification and clustering both involve categorization of objects. Classification processes a previously known categorization of objects from a training sample so that it can be applied to other objects whose categorization is unknown. This process is called *supervised learning*. Clustering groups objects with similar characteristics. As opposed to classification, the grouping process in clustering is unsupervised. The actual categorization of objects, even for a sample, is unknown. Clustering is an important step in establishing object profiles.

Clustering in data mining faces several additional challenges compared to conventional clustering applications (Krishnapuram, Joshi, Nasraoui, & Yi, 2001). The clusters tend to have fuzzy boundaries. There is a likelihood that an object may be a candidate for more than one cluster. Krishnapuram et al. argued that clustering operations in data mining involve modeling an unknown number of overlapping sets. This article describes fuzzy and interval set clustering as alternatives to conventional crisp clustering.

BACKGROUND

Conventional clustering assigns various objects to precisely one cluster. Figure 1 shows a possible clustering of 12 objects. In order to assign all the objects to precisely one cluster, we had to assign objects 4 and 9 to Cluster B. However, the dotted circle seems to represent a more natural representation of Cluster B. An ability to specify that Object 9 may either belong to Cluster B or Cluster C, and Object 4 may belong to Cluster A or Cluster B, will provide a better representation of the clustering of the 12 objects in Figure 1. Rough-set theory provides such an ability.

The notion of rough sets was proposed by Pawlak (1992). Let U denote the universe (a finite ordinary set), and let R be an equivalence (indiscernibility) relation on U. The pair $A=(U,R)$ is called an *approximation space*. The equivalence relation R partitions the set U into disjoint subsets. Such a partition of the universe is denoted by $U/R=\{E_1,E_2,\cdots,E_m\}$, where E_i is an equivalence class of R. If two elements $\mathbf{u},\mathbf{v}\in U$ belong to the same equivalence class, we say that **u** and **v** are indistinguishable. The equivalence classes of R are called the *elementary* or *atomic sets* in the approximation space $A=(U,R)$. Because it is not possible to differentiate the elements within the same equivalence class, one may not be able to obtain a precise representation for an arbitrary set $X\subseteq U$ in terms of elementary sets in A. Instead, its lower and upper bounds may represent the set X. The lower bound $\underline{A}(X)$ is the union of all the elementary sets, which are subsets of X. The upper bound $\overline{A}(X)$ is the union of all the elementary sets that have a nonempty intersection with X. The pair $(\underline{A}(X),\overline{A}(X))$ is the representation of an ordinary set X in the approximation space $A=(U,R)$, or simply the rough set of X. The elements in the lower bound of X definitely belong to X, while elements in the upper bound of X may or may not belong to X. The pair

Figure 1. Conventional clustering

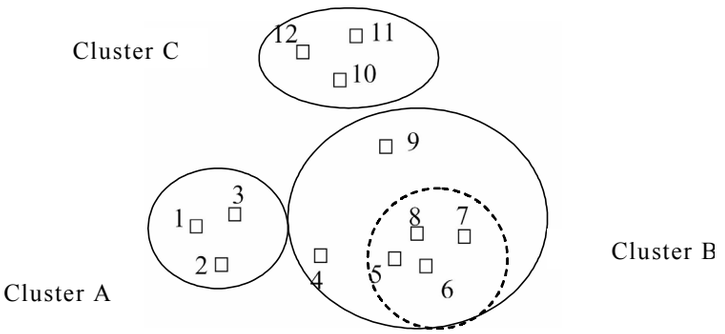
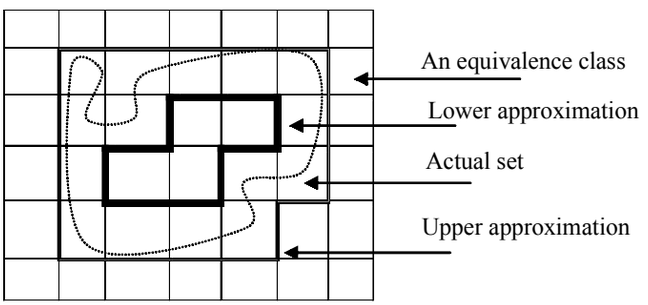


Figure 2. Rough set approximation



$(\underline{A}(X), \overline{A}(X))$ also provides a set theoretic interval for the set X. Figure 2 illustrates the lower and upper approximation of a set.

MAIN THRUST

Interval Set Clustering

Rough sets were originally used for supervised learning. There are an increasing number of research efforts on clustering in relation to rough-set theory (do Prado, Engel, & Filho, 2002; Hirano & Tsumoto, 2003; Peters, Skowron, Suraj, Rzasa, & Borkowski, 2002). Lingras (2001) developed rough-set representation of clusters. Figure 3 shows how the 12 objects from Figure 1 could be clustered by using rough sets. Instead of Object 9 belonging to any one cluster, it belongs to the upper bounds of Clusters B and C. Similarly, Object 4 belongs to the upper bounds of Clusters A and B.

Lingras (2001; Lingras & West, 2004; Lingras, Hogo, & Snorek, 2004) proposed three different approaches for unsupervised creation of rough or interval set representations of clusters: evolutionary, statistical, and neural. Lingras (2001) described how a rough-set theoretic clustering scheme could be represented by using a rough-set genome. The rough-set genome consists of one gene per object. The gene for an object is a string of bits that describes which lower and upper approximations the object belongs to. The string for a gene can be partitioned into two parts, lower and upper, as shown in Figure 4 for three clusters. Both lower and upper parts of the string consist of three bits each. The i^{th} bit in the lower/upper string tells whether the object is in the lower/upper approximation of the i^{th} cluster. Figure 4 shows examples of all the valid genes for three clusters. An object represented by g_1 belongs to the upper bounds of the first and second clusters. An object represented by g_6 belongs to the lower and upper bounds of the second cluster. Any other value not given by g_1 to g_7 is not valid. The objective of the genetic algorithms (GAs) is to minimize the within-group-error. Lingras provided a formulation of within-group-error for rough-set based clustering. The resulting GAs were used to evolve interval clustering of highway sections. Lingras (2002) applied the unsupervised rough-set clustering based on GAs for grouping Web users. However, the clustering process based on GAs seemed computationally expensive for scaling to larger datasets.

The *K*-means algorithm is one of the most popular statistical techniques for conventional clustering (Hartigan & Wong, 1979). Lingras and West (2004) provided a theoretical and experimental analysis of a modified *K*-means clustering based on the properties of rough sets. It was used to create interval set representa-