Chapter 12
Data Clustering Algorithms Using Rough Sets

B.K.Tripathy
VIT University, India

Adhir Ghosh
VIT University, India

ABSTRACT

Developing Data Clustering algorithms have been pursued by researchers since the introduction of k-means algorithm (Macqueen 1967; Lloyd 1982). These algorithms were subsequently modified to handle categorical data. In order to handle the situations where objects can have memberships in multiple clusters, fuzzy clustering and rough clustering methods were introduced (Lingras et al 2003, 2004a). There are many extensions of these initial algorithms (Lingras et al 2004b; Lingras 2007; Mitra 2004; Peters 2006, 2007). The MMR algorithm (Parmar et al 2007), its extensions (Tripathy et al 2009, 2011a, 2011b) and the MADE algorithm (Herawan et al 2010) use rough set techniques for clustering. In this chapter, the authors focus on rough set based clustering algorithms and provide a comparative study of all the fuzzy set based and rough set based clustering algorithms in terms of their efficiency. They also present problems for future studies in the direction of the topics covered.

INTRODUCTION

A cluster is a collection of data objects that are similar to one another. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. So, it has largely been used as a data analysis tool to characterize data sets. It has been used in data mining tasks such as unsupervised classification and data summation, as well as segmentation of large heterogeneous data sets into smaller homogeneous subsets that can be easily managed, separately modeled and analyzed (Huang 1998). Cluster Analysis has been widely used in numerous applications, including market research, pattern recognition, image processing, research and development, nuclear science, medicine and in business. In business, for example clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on their purchasing patterns.

There are several applications of cluster analysis. We only name a few of them here. Jiang et al (2004) analyze a variety of cluster techniques for

DOI: 10.4018/978-1-4666-2518-1.ch012

complex gene expression data. Wu et al (2004) have developed a clustering algorithm specifically designed to handle the complexities of gene data that can estimate the correct number of clusters and find them. Wong et al (2002) present an approach used to segment tissues in a nuclear medical imaging method known as positron emission tomography (PET). Mathieu and Gibson (2004) use cluster analysis as a part of a decision support tool for large-scale research and development planning to identify programs to participate in and to determine resource allocation. Haimov et al (1989) use cluster analysis to segment radar signals in scanning land and marine objects. Saglam et al. expressed the clustering problem in the form of a mixed-integer programming problem with the objective of minimizing the maximum cluster diameter among all clusters. This was applied to solve the customer segmentation problem of a digital platform company involving demographic and transactional attributes related to the customers. Fathian et al proposed a hybridization of nature inspired intelligent technique with K-means algorithm. Chen and Liu (2009) proposed an effective clustering algorithm, which was used to resolve the classification problem of construction management.

Cluster analysis is a challenging field of research as it is applied in several diverse fields. Clustering is sometimes called data segmentation as it divides a data set into several groups depending upon the similarity of individual elements. In data mining, clustering needs to possess certain characteristics like scalability, handling of hybrid data, generating clusters with random shape, handling missing values, parameter identification, handling of dynamic updation of data values and dealing with large number of attributes.

In conventional clustering the data with similar characteristics are grouped together to form a single cluster However, in practice it has been observed that this requirement is very stringent and objects may show characteristics to belong to several clusters. In such cases an object may belong to more than one clusters leading to overlapping of clusters instead of them being distinct. In order to handle such situations multiple memberships of objects became a necessity. This led to the development of clustering algorithms using fuzzy techniques. In later developments rough set techniques were used in developing such algorithms, which also handles uncertainty of data in an efficient manner. Rough set based clustering provides a solution that is less restrictive than conventional clustering and less descriptive than fuzzy clustering (Lingras et al 2003).

In this chapter, we discuss briefly on the chronological development of different clustering algorithms, starting from conventional to fuzzy based ones. The primary focus of the chapter is to discuss the rough set based algorithms in detail, provide a comparative analysis of these algorithms and compare their efficiency with other fuzzy based algorithms. Also, we provide some directions of research for further study.

The organization of the chapter is as follows. In the next section, we provide a background of different clustering algorithms. Section 3 deals with the mathematical back ground of the topics to be covered in this chapter. In section 4 we present the most recent algorithms developed using rough set theory. In section 5 we provide a comparative study of the three similar type algorithms, MMR, MMeR and SDR with respect to different characteristics. In section 6 we present an empirical study of all the algorithms handling uncertainty by taking different data sets available in the UCI data repository. Section 7 deals with some future directions of research on the topic dealt with in the chapter. We end up with an exhaustive reference of materials consulted during the compilation of the chapter.

BACKGROUND

In this section we provide the origin and development of the clustering algorithms in the literature.