Chapter LX
Data Clustering

Yanchang Zhao
University of Technology, Sydney, Australia

Longbing Cao
University of Technology, Sydney, Australia

Huaifeng Zhang
University of Technology, Sydney, Australia

Chengqi Zhang
University of Technology, Sydney, Australia

INTRODUCTION

Clustering is one of the most important techniques in data mining. This chapter presents a survey of popular approaches for data clustering, including well-known clustering techniques, such as partitioning clustering, hierarchical clustering, density-based clustering and grid-based clustering, and recent advances in clustering, such as subspace clustering, text clustering and data stream clustering. The major challenges and future trends of data clustering will also be introduced in this chapter.

The remainder of this chapter is organized as follows. The background of data clustering will be introduced in Section 2, including the definition of clustering, categories of clustering techniques, features of good clustering algorithms, and the validation of clustering. Section 3 will present main approaches for clustering, which range from the classic partitioning and hierarchical clustering to recent approaches of bi-clustering and semi-supervised clustering. Challenges and future trends will be discussed in Section 4, followed by the conclusions in the last section.

BACKGROUND

Data clustering is sourced from pattern recognition (Theodoridis & Koutroumbas, 2006), machine learning (Alpaydin, 2004), statistics (Hill & Lewicki, 2007) and database technology (Date, 2003). *Data clustering* is to partition data into groups, where the data in the same group are similar to one another and the data from different groups are dissimilar (Han & Kamber, 2000). More specifically, it is to segment data into clusters

so that the intra-cluster similarity is maximized and that the inter-cluster similarity is minimized. The groups obtained are a partition of data, which can be used for customer segmentation, document categorization, etc.

Clustering techniques can be "clustered" into groups in multiple ways. In terms of the membership of objects, there are two kinds of clustering, *fuzzy clustering* and *hard clustering*. Fuzzy clustering is also known as *soft clustering*, where an object can be in more than one cluster, but with different membership degrees. In contrast, an object in hard clustering can belong to one cluster only. Generally speaking, clustering is referred to as hard clustering implicitly. In terms of approaches, data clustering techniques can be classified into the following groups: partitioning clustering, hierarchical clustering, density-based clustering, grid-based clustering and model-based clustering. In terms of the type of data, there are spatial data clustering, text clustering, multimedia clustering, time series clustering, data stream clustering and graph clustering.

For a good clustering algorithm, it is supposed to have the following features: 1) the ability to detect clusters with various shapes and different distributions; 2) the capability of finding clusters with considerably different sizes; 3) the ability to work when outliers are present; 4) no or few parameters needed as input; and 5) scalability to both the size and the dimensionality of data.

How to evaluate the results is an important problem for clustering. For the validation of clustering results, there are many different measures, such as *Compactness* (Zait & Messatfa, 1997), *Conditional Entropy (CE)* and *Normalized Mutual Information (NMI)* (Strehl & Ghosh, 2002; Fern & Brodley, 2003). The validation measures can be classified into three categories, 1) *internal validation*, such as *Compactness*, *Dunn's validation index*, *Silhouette index* and *Hubert's correlation with distance matrix*, which is based on calculating the properties of result clusters, 2) *relative validation*, such as *Figure of merit* and *Stability*, which is based on comparisons of partitions, and 3) *external validation*, such as *CE*, *NMI*, *Hubert's correlation*, *Rand statistics*, *Jaccard coefficient*, and *Folkes and Mallows index*, which is based on comparing with a known true partition of data (Halkidi et al., 2001, Brun et al., 2007).

DATA CLUSTERING TECHNIQUES

The popular clustering techniques will be briefly presented in this section. More detailed introduction and comparison of various clustering techniques can be found in books on data mining and survey papers on clustering (Berkhin, 2002; Grabmeier & Rudolph, 2002; Han & Kamber, 2000; Jain, Murty, & Flynn, 1999; Kolatch, 2001; Xu & Wunsch, 2005; Zait & Messatfa, 1997).

Partitioning Clustering

The idea of *partitioning clustering* is to partition the data into *k* groups first and then try to improve the quality of clustering by moving objects from one group to another. A typical method of partitioning clustering is *k-means* (Alsabti, Ranka, & Singh, 1998; Macqueen, 1967), which randomly selects *k* objects as cluster centers and assigns other objects to the nearest cluster centers, and then improves the clustering by iteratively updating the cluster centers and reassigning the objects to the new centers. *k-medoids* (Huang, 1998) is a variation of *k*-means for categorical data, where the medoid (i.e., the object closest to the center), instead of the centroid, is used to represent a cluster. Some other partitioning methods are PAM and CLARA proposed by Kaufman & Rousseeuw (1990) and CLARANS by Ng and Han (1994).

The disadvantage of partitioning clustering is that the result of clustering is dependent on the selection of initial cluster centers and it may result in a local optimum instead of a global one. A simple way to improve the chance of obtaining the global optimum is to run *k*-means