Chapter LII
Categorization of Data Clustering Techniques

Baoying Wang
Waynesburg University, USA

Imad Rahal
College of Saint Benedict, Saint John's University, USA

Richard Leipold
Waynesburg University, USA

INTRODUCTION

Data clustering is a discovery process that partitions a data set into groups (clusters) such that data points within the same group have high similarity while being very dissimilar to points in other groups (Han & Kamber, 2001). The ultimate goal of data clustering is to discover natural groupings in a set of patterns, points, or objects without prior knowledge of any class labels. In fact, in the machine-learning literature, data clustering is typically regarded as a form of unsupervised learning as opposed to supervised learning. In unsupervised learning or clustering, there is no training function as in supervised learning. There are many applications for data clustering including, but not limited to, pattern recognition, data analysis, data compression, image processing, understanding genomic data, and market-basket research.

BACKGROUND

Data clustering is an important human activity. As humans, we can easily perform mental tasks such as distinguishing between cats and dogs, or between animals and plants. A more concrete example of clustering is given in Figure 1, which demonstrates the clustering of padlocks. In this example, there are 10 padlocks with different colors and shapes that we would like to cluster into three different groups.

Figure 1. A simple example of clustering

(a) The padlocks before clustering

(b) The padlocks after clustering

Clustering has its roots in a number of fields including data mining, statistics, biology, and machine learning. The importance and interdisciplinary nature of clustering is evident in its rich and diverse literature. Besides the phrase data clustering, a number of other terms and phrases have been coined to describe the same process, namely, cluster analysis, automatic classification, numerical taxonomy, botryology, and typological analysis (Jain & Dubes, 1988).

Representing data by fewer clusters necessarily introduces a loss of certain fine details such as specific properties pertaining to individual data objects, but achieves the more important goal of simplification—a highly desirable characteristic in an age of inexorable abundance of data. Clustering represents many data objects using a few clusters; hence, it models the data by these clusters. From a machine-learning perspective, the clusters correspond to hidden patterns where the search for clusters can be viewed as a form of unsupervised learning and the resulting system as a representation of a data concept. Consequently, clustering is the unsupervised learning of a hidden data concept.

Data mining focuses primarily on large databases that impose additional severe computational requirements on data clustering as a process. These challenges led to the emergence of numerous powerful and broadly applicable clustering approaches (Berkhin, 2002).

DATA CLUSTERING TECHNIQUES

Categorization of Data Clustering Techniques

Generally speaking, data clustering techniques can be categorized in a number of distinct ways (Berkhin, 2002; Han & Kamber, 2001; Jain & Dubes, 1988), one of which, based on the structure of the produced clusters, is shown in Figure 2.

As depicted, clustering can be subdivided into partitioning clustering, hierarchical clustering, and hybrid clustering. Hierarchical clustering produces a nested sequence of partitions, whereas partitioning clustering results in a single partition. Hierarchical clustering approaches can be further categorized into agglomerative (bottom-up) and divisive (top-down) hierarchical clustering depending on whether the hierarchical decomposition is carried in a bottom-up or a top-down fashion. Partitioning clustering consists of two subcategories, distance based and density based, depending on the similarity measure utilized by the clustering process. A recent trend has been to combine features of hierarchical and partitioning