Chapter 9
Novel Trends in Clustering

Claudia Plant
Technische Universität München, Munich Germany, Ludwig Maximilians Universität München, Munich, Germany

Christian Böhm
Technische Universität München, Munich Germany, Ludwig Maximilians Universität München, Munich, Germany

ABSTRACT

Clustering or finding a natural grouping of a data set is essential for knowledge discovery in many applications. This chapter provides an overview on emerging trends within the vital research area of clustering including subspace and projected clustering, correlation clustering, semi-supervised clustering, spectral clustering and parameter-free clustering. To raise the awareness of the reader for the challenges associated with clustering, the chapter first provides a general problem specification and introduces basic clustering paradigms. The requirements from concrete example applications in life sciences and the web provide the motivation for the discussion of novel approaches to clustering. Thus, this chapter is intended to appeal to all those interested in the state-of-the art in clustering including basic researchers as well as practitioners.

INTRODUCTION

In many applications, for example in medicine, life sciences, physics and market observation, terabytes of data is collected every day. Consider for example metabolite profiling (Baumgartner & Graber 2008). As an evolving branch of life sciences, Metabolomics studies the highly complex metabolism of cells, tissues, organs and organisms. One major focus of research is on identifying subtle changes related to disease onset and progression. Small molecules involved in primary and intermediate metabolism are called metabolites. Metabolite profiling provides techniques to quantify the amount of metabolites in a sample. Due to recent advances of high-throughput technologies such as tandem mass spectrometry (MS/MS) hundreds of metabolites can be detected from a single blood sample. As a second example consider web usage. For each user accessing a page, the corresponding web server logs information including IP address, time of access, file path, browser and amount of transferred data.

DOI: 10.4018/978-1-60566-816-1.ch009

Huge volumes of web server log data is generated every day and its potential for commercial and non-commercial applications such as designing online shops or providing users with personalized content in digital libraries (Zaiane & al. 1998) is far from being fully exploited.

In both applications scenarios, extraction of information from the massive amounts of data is a non-trivial, highly challenging task. In both scenarios we want to learn unknown regularities and structure in the data with very little previous knowledge. In metabolite profiling, we want to gain novel insights how certain diseases change the pattern of metabolites. Simple statistic tests often applied in biomedicine can provide valuable information. However, only a tiny part of the information potentially available in the data can be accessed but large parts remain unexplored. There may be several sub-types of the disease each associated with a unique pattern of altered metabolism. Also in the healthy controls there may be different types of normal yet unexplored metabolic patterns. Similarly, in the second scenario we want to find groups of users with similar behavior to provide them personalized content.

As an important area within data mining, clustering aims at partitioning the data into groups such that the data objects assigned to a common group called cluster are as similar as possible and the objects assigned to different clusters differ as much as possible. With the term 'data objects' we denote the instances subjected to a cluster analysis. Often, data objects can be represented as a feature vectors. In the scenario of metabolite profiling, the data objects are the subjects. Each subject is represented by a vector composed of the amounts of the measured metabolites. The dimensionality of the resulting feature space equals the number of metabolites. Alternatively, it could also be interesting to cluster the metabolites in the space defined by the subjects with the objective to identify groups of metabolites having similar prevalence across subjects.

Figure 1 displays examples of different types of clusters in vector data. The simplest type of a cluster is a spherical Gaussian. An example in two-dimensional space is depicted in Figure 1(a). Both coordinates follow a Gaussian distribution and are statistically independent from each other. As we will see in the next section, basic clustering algorithms can reliably detect such clusters. More complicated are correlation clusters with orthogonal major directions, as depicted in Figure 1(b). The objects of this cluster follow a line in one-dimensional space which is characterized by a strong linear dependency between the coordinates. In addition, the major directions of the cluster are orthogonal and can be detected by Principal Component Analysis. Figure 1(c) displays a non-linear correlation cluster. There exists a distinct dependency between the two coordinates but this dependency cannot be captured by a linear

Figure 1. Different types of clusters in vector data: (a) spherical Gaussian; (b) correlation cluster; (c) non-linear correlation cluster; (d) non-Gaussian cluster with non-orthogonal major directions