Visualizing Cluster Analysis and Finite Mixture Models

III.11

Friedrich Leisch

11.1 Introduction

Data visualization can greatly enhance our understanding of multivariate data structures, and so it is no surprise that cluster analysis and data visualization often go hand in hand, and that textbooks like Gordon (1999) or Everitt et al. (2001) are full of figures. In particular, hierarchical cluster analysis is almost always accompanied by a dendrogram. Results from partitioning cluster analysis can be visualized by projecting the data into two-dimensional space or using parallel coordinates. Cluster membership is usually represented by different colors and glyphs, or by dividing clusters into several panels of a trellis display (Becker et al., 1996). In addition, silhouette plots (Rousseeuw, 1987) provide a popular tool for diagnosing the quality of a partition. Some of the popularity of self-organizing feature maps (Kohonen, 1989) with practitioners in various fields can be explained by the fact that the results can be "easily" visualized.

In this chapter we provide an overview of visualization techniques for cluster analysis results. Using two real-world data sets, we explain the most important types of graphs that can be used in combination with hierarchical, partitioning and model-based cluster analysis. Many plots like dendrograms, convex cluster hulls or silhouettes are specific to clustering, but we also demonstrate how graphical techniques introduced in other chapters of this handbook can be used as building blocks for cluster visualization.

11.1.1 The Data Sets

Two data sets are used throughout this chapter. The "dentitio" data set is used for hierarchical clustering (e.g., Hartigan, 1975). This data set gives the counts for eight kind of teeth – top-jaw and bottom-jaw counts for incisors, canines, premolars and molars – in 66 different species of animals. A subset of the raw data is listed in Table 11.1.

The second data set, which is used for partitioning and model-based clustering in Sects. 11.3 and 11.4, is related to the German parliamentary elections of September 18, 2005. A subset of the raw data is given in Table 11.2. The data consist of the proportions of the "second votes" obtained by the five parties that got elected to the Bundestag (the first chamber of the German parliament) for each of the 299 electoral districts. The "second votes" are actually more important than the "first votes" because they control the number of seats each party has in parliament. Note that the proportions do not sum to 1 because parties that did not get elected into parliament have been omitted from the table.

Before election day, the German government comprised a coalition of Social Democrats (SPD) and the Green Party (GRUENE); their main opposition consisted of the conservative party (Christian Democrats, CDU/CSU) and the Liberal Party (FDP). The latter two intended to form a coalition after the election if they gained a joint majority, so the two major "sides" during the campaign were SPD+GRUENE versus CDU/CSU+FDP. In addition, a new "party of the left" (LINKE) canvassed for