Chapter 2

Getting to Know Your Data

2.1 Bibliographic Notes

Methods for descriptive data summarization have been studied in the statistics literature long before the onset of computers. Good summaries of statistical descriptive data mining methods include Freedman, Pisani and Purves [FPP07], and Devore [Dev95]. For statistics-based visualization of data using boxplots, quantile plots, quantile-quantile plots, scatter plots, and loess curves, see Cleveland [Cle93].


There have been many graphical user interfaces and visualization tools developed for data mining, which can be found in various data mining products. Several books on data mining, such as Data Mining Solutions by Westphal and Blaxton [WB98], present many good examples and visual snapshots. For a survey of visualization techniques, see “Visual techniques for exploring databases” by Keim [Kei97].

Similarity and distance measures among various variables have been introduced in many textbooks that study cluster analysis, including Hartigan [Har75], Jain and Dubes [JD88], Kaufman and Rousseeuw [KR90], and Arabie, Hubert, and De Sorte [AHS96]. Methods for combining attributes of differ-
ent types into a single dissimilarity matrix were introduced by Kaufman and Rousseeuw [KR90].
Bibliography


