

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].

Pioneering work on data visualization techniques is described in *The Visual Display of Quantitative Information* [Tuf83], *Envisioning Information* [Tuf90], and *Visual Explanations : Images and Quantities, Evidence and Narrative* [Tuf97], all by Tufte, in addition to *Graphics and Graphic Information Processing* by Bertin [Ber81], *Visualizing Data* by Cleveland [Cle93], and *Information Visualization in Data Mining and Knowledge Discovery* edited by Fayyad, Grinstein, and Wierse [FGW01]. Major conferences and symposiums on visualization include *ACM Human Factors in Computing Systems (CHI)*, *Visualization*, and the *International Symposium on Information Visualization*. Research on visualization is also published in *Transactions on Visualization and Computer Graphics*, *Journal of Computational and Graphical Statistics*, and *IEEE Computer Graphics and Applications*.

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

- [AHS96] P. Arabie, L. J. Hubert, and G. De Soete. *Clustering and Classification*. World Scientific, 1996.
- [Ber81] J. Bertin. *Graphics and Graphic Information Processing*. Berlin, 1981.
- [Cle93] W. Cleveland. *Visualizing Data*. Hobart Press, 1993.
- [Dev95] J. L. Devore. *Probability and Statistics for Engineering and the Sciences* (4th ed.). Duxbury Press, 1995.
- [FGW01] U. Fayyad, G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann, 2001.
- [FPP07] D. Freedman, R. Pisani, and R. Purves. *Statistics (4th ed.)*. W. W. Norton & Co., 2007.
- [Har75] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, 1975.
- [JD88] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.
- [Kei97] D. A. Keim. Visual techniques for exploring databases. In *Tutorial Notes, 3rd Int. Conf. Knowledge Discovery and Data Mining (KDD'97)*, Newport Beach, CA, Aug. 1997.
- [KR90] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.
- [Tuf83] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.
- [Tuf90] E. R. Tufte. *Envisioning Information*. Graphics Press, 1990.
- [Tuf97] E. R. Tufte. *Visual Explanations : Images and Quantities, Evidence and Narrative*. Graphics Press, 1997.
- [WB98] C. Westphal and T. Blaxton. *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*. John Wiley & Sons, 1998.