Book Review

Data Mining: Concepts, Models, Methods and Algorithms Mehmood Kantardzic (New York: John Wiley and IEEE, 2003, ISBN 0-471-22 852-4. *Reviewed by Jacek M. Zurada*

The importance of data mining arises from the fact that the modern world is increasingly data-driven. We are surrounded by data in numerical, symbolic, verbal and visual formats, to name a few. The data, often raw, must be analyzed and processed to convert it into other formats that inform, instruct, answer, or otherwise aid understanding and assist us in decision-making. In the age of Internet, intranets, data warehouses and data marts, available data sets have grown in size and complexity. This, in turn, has lead to an inevitable shift away from direct hands-on data analysis toward indirect, automatic and intelligent data analysis in which the analyst works via increasingly complex and sophisticated software tools.

This growth has also brought the emergence of a new area of interest in data mining, developed especially to extract valuable information from very large data sets. Short of being a new discipline equipped with a new core of knowledge, data mining offers an interdisciplinary collection of specialized techniques that allow for integration and enhancement of results from disciplines such as statistics, artificial intelligence, data bases, pattern recognition, and computer visualization.

This new book provides a systematic review of state-of-the-art methodologies and techniques for analyzing large quantities of raw data in high-dimensional data spaces, and for extraction of new information for decision-making. While data mining has its roots in several disciplines, most existing books have been biased with background and emphasis in one of the fields such as data bases, statistics, pattern recognition, artificial intelligence, or business information systems [1]–[8]. This book is trying to present a comprehensive and balanced approach by including most of the relevant results in databases, machine learning, statistics, computer visualization, and soft computing that are contributing to the data mining field. The scope of the book is somewhat similar to Berthold's and Hand's volume [9], yet the level of presentation makes it appealing to the readers that have less rigorous mathematical background.

The book covers a wide range of current theoretical and practical issues. The emphasis of the presentation is on detailed explanations of relevant phases and techniques of the data mining process. These are followed by illustrative, often numerical, examples. The lucent approach is likely to widen the readership for this book as compared with many other books that require stronger mathematics or computer science prerequisites. Although the book is introductory in nature, elementary background in algorithms, statistics, data structures, and databases will be helpful.

After introducing introductory concepts of data mining in Chapter 1, Chapters 2–3 cover common characteristics of raw large data sets and typical techniques of data preprocessing. Topics covered include transformations, handling of missing data and outliers, reduction of features and data sets, entropy, and principal component analysis, to name a few. The text emphasizes the importance of these initial phases of transforming and reducing large raw data sets for the final success and quality of data mining results.

Chapter 4 introduces the theoretical framework of learning in observational environment. Chapters 5–11 provide an overview of commonly used approaches of data mining techniques such as statistical inference, clustering, logic-based techniques, association rules, web and text mining techniques, artificial neural networks, genetic algorithms, and fuzzy systems. Chapter 12 reviews a number of visualization techniques especially useful for representation of high-dimensional samples. Appendixes provide an overview of commercially and publicly available data mining packages and offer an extensive address list of relevant websites and software vendors.

By timely publishing of this book, Dr. Kantardzic has provided the technical community with an informative and very readable volume in this fast growing area. The volume is also a very good candidate for a textbook in an introductory data mining course. Not only is it organized in a systematic, pedagogical way, but also provides a number of Review Questions and Problems at the end of each chapter. These should be very helpful for adopters and students alike. In fact, this book may encourage instructors to develop a new course in data mining at the upper undergraduate or at the entry graduate level in engineering and computer science programs.

References

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Diego, CA: Academic, 2001.
[2] D. Hand, H. Mannila, and P. Smith, *Principles of Data Mining*. Cambridge, MA: MIT Press, 2001.
[3] V. Cherkassky and F. Mulier, *Learning From Data: Concepts, Theory and Methods*. New York: Wiley, 1998.
[4] A. Berson, S. Smith, and K. Thearling, *Building Data Mining Applications for CRM*. New York: McGraw-Hill, 2000.
[5] C. Westphal and T. Blaxton, *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*. New York: Wiley, 1998.
[6] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations*. San Francisco, CA: Morgan Kaufman, 1999.
[7] R. L. Kennedy, Y. Lee, B. Van Roy, C. D. Reed, and R. P. Lippman, *Solving Data Mining Problems Through Pattern Recognition*. Upper Saddle River, NJ: Prentice-Hall, 1998.
[8] R. Groth, *Data Mining: A Hands-On Approach for Business Professionals*. Upper Saddle River, NJ: Prentice-Hall, 1998.
[9] *Intelligent Data Analysis—An Introduction*, M. Berthold and D. Hand, Eds., Springer-Verlag, Berlin, Germany, 1999.

Digital Object Identifier 10.1109/TNN.2003.809639