



# NEW BOOKS IN REVIEW

EDITOR: *George R. Franke*

ASSOCIATE EDITORS: *Naveen Donthu*

*Meryl P. Gardner*

HANDBOOK OF DATA MINING AND KNOWLEDGE DISCOVERY, Willi Klösgen and Jan M. Żytkow, eds., Oxford: Oxford University Press, 2002, 1026 pages, \$250.00.

Data mining, a collection of analytic methods for extracting information from large databases, has had a difficult time gaining acceptance among marketing academics. Aside from the term's pejorative origins of data dredging or fishing, data mining clashes with the modeling approaches that prevail in marketing journals. Marketing academics are reared in an analytic framework that cherishes theory-based models that can be validated with data. From a theory-testing perspective, data mining is of questionable value. However, from a different point of view, one that values predictive accuracy above all else, data mining is a marvel. Data mining is about algorithms that generate accurate predictions; it is not about models that validate theories. Theory testing is usually performed on static and self-contained data sets, many of which are assembled for the particular theoretical question at hand (e.g., collecting preference rankings to evaluate the importance of attributes with conjoint analysis). Data mining is performed on large data sets that are constantly updated with new information. This scenario plays itself out every day in many areas of marketing practice, including retailing, financial services, and direct marketing.

Data miners concern themselves with distilling meaningful information and predictions from ever increasing volumes of data. Marketing researchers should not dismiss data mining because it lacks a theoretical base any more than data miners should attack theoretical modeling for its failure to represent patterns in massive amounts of data. Largely because of the incongruence between data mining and traditional methods of analysis in marketing, the academic field of marketing has been left behind in the development of data mining. A diverse set of fields, such as computational science, engineering, and artificial intelligence, has come together to develop analytic techniques, data management systems, software, and a multitude of applications. Oddly enough, although the marketing journals have had little to say about data mining, marketing practice is one of the most frequently studied areas in data mining literature.

*Handbook of Data Mining and Knowledge Discovery* surveys a broad set of topics that make up data mining and is a good reference for academic researchers considering these methods, researchers already familiar with some data min-

ing techniques, teachers of database marketing courses, and serious practitioners looking for additional background. The term "data mining" is sometimes synonymous with the term "knowledge discovery in databases" (KDD), though some authors use data mining to mean the application of algorithms to find patterns and use KDD to include also the steps before and after applying the algorithms. The preface states (p. xxii), "the handbook provides ... an overview of the field, collating and filtering the research findings into a sometimes less detailed but broader view of the domain, and satisfies the need for a broad-based reference book." Furthermore (p. xxi), "Since KDD has emerged by the creative combination of so many fields, even an expert in one field lacks experience in many others, so that a broad summary of contributions from all the founding fields will facilitate interdisciplinary understanding and communication.... The handbook provides fast access to the basic concepts." Every chapter has extensive references to academic articles, books, and Web sites for those who are interested in further reading. In the chapters on unfamiliar topics, we found the problems being solved and the gist of how the methods work to be reasonably understandable. The book is scholarly and free of the hype often found in trade-press articles on data mining. It focuses on data mining methods in general rather than how data mining is used or could be used in marketing; readers must make this connection themselves.

The people who develop, study, and use data mining methods are a heterogeneous group. As David Hand notes (p. 638) in Chapter 25, "data mining is an eclectic discipline. It has taken ideas and tools from a wide variety of areas to apply to the particular kind of problems its practitioners wish to solve." To capture this diversity, more than 100 researchers and practitioners contributed entries to the *Handbook*. This approach avoids a limitation of many other data mining books, which often present only one perspective on data mining. The danger of having so many contributors is that the entries can be inconsistent in style, scope, rigor, and terminology; however, this book seems to have been carefully edited, and such inconsistencies are not pronounced.

What is data mining? Several chapters address this question. Hand defines (p. 637) data mining as "the process or secondary analysis of large databases aimed at finding unsuspected relationships that are of interest or value to the database owners." He elaborates on specific terms in this definition and juxtaposes data mining with more traditional statistical analyses. To a data miner, "large databases"

means databases whose size is measured in terabytes. These databases often grow over time; in marketing, additional transactions are continually added to point-of-sales databases, credit card databases, telecommunications databases, and so on. Data mining develops fast, scalable algorithms that can be applied in these situations. *Scalable* means that the algorithm remains computationally feasible as the size of the data set increases. Many algorithms are *adaptive*, which means that the estimates from the methods can dynamically change as more data become available. As Hand notes (p. 640), the need for real-time responses informed by analyses of massive data sets has led data miners to prefer computationally simple models for many problems. Compare this with the computationally demanding models often found in marketing journals.

Hand (see also Breiman 2001; Hand 1998) distinguishes between the statistical models used in much academic research, including marketing, and the algorithmic models developed in data mining. Breiman (1994, pp. 38–39) captures this difference: “[T]here are important cultural differences between the statistical and neural network communities. If a statistician analyzes data, the first question he gets asked is ‘what’s your data model?’ The [data mining] practitioner will be asked ‘what’s your accuracy?’” To data miners, dependent variables are related to independent variables by some black box. A common data mining goal is to model the relationship as accurately as possible. Contrast this with the statistical modeling approach, which attempts to propose “a stochastic data model for the inside of the black box” (Breiman 2001, p. 199). The theory explaining how and why inputs are related to outputs is of secondary importance to a data miner, just as predictive accuracy is of secondary importance to a statistical modeler. A focus on predictive accuracy makes perfect sense for machine-learning problems such as handwriting recognition, fingerprint identification, and classification of tumors on a mammogram. In marketing, it makes sense for operational tasks such as predicting the probability that someone will respond to a direct-mail offer or determining which three books Amazon.com should cross-sell someone who searches for a particular book. However, is optimizing predictive accuracy important when the objective is, for example, to develop consumer behavior theory or a high-level marketing strategy?

#### *What Will JMR Readers Find Interesting?*

The book is organized into eight parts, each with multiple chapters and sections. Individual sections are written by experts on the topic. Part 1 motivates the need for these methods and attempts to define data mining and KDD. It describes how data mining has emerged from fields that include logic, statistics, database, artificial intelligence, automated scientific discovery, neural networks, machine learning, pattern recognition, and data visualization. Part 2 discusses fundamental concepts, including primers on types of data, database systems, logic, probability and statistics, rough sets, fuzzy sets, and search techniques. Many of these chapters will be of interest to *JMR* readers. For example, as teachers of database marketing courses, we found the primer on database systems helpful. It discusses the characteristics and uses of various types of database systems, including relational databases, object-oriented databases, multidimen-

sional databases/online analytical processing, and parallel databases. Because these systems enable data-driven marketing programs, it is beneficial to have some background.

Part 3 discusses the individual tasks involved in data mining. The first task usually involves interacting with a data warehouse; the *Handbook* provides useful discussion on practical aspects such as data cleaning, quality assurance, and security. These are topics about which anyone teaching database marketing topics to MBA students or executives should know. Doctoral students and academic researchers doing empirical research with real-world data will increasingly find these methods useful. The *Handbook* also provides overviews of a broad range of data mining methods for classification (categorical dependent variables), regression (numerical dependent variables), clustering, association rules/market basket analysis, and other problems. It also has sections on scalability and parallel computation methods, topics that marketing methodologists will need to consider as the size of marketing databases grows.

Many sections are written by researchers who have made seminal contributions on their subject. The decision-tree section is coauthored by J. Ross Quinlan, who has developed one of the leading implementations, C5.0, and has written many articles and books on the subject. Heikki Mannila, who coauthored some of the seminal papers on the a priori algorithm used in market-basket analysis, wrote the section on association rules. Readers of *JMR* will find some of the vernacular strange; for example, the “regression” chapter focuses mostly on linear discriminant analysis, and “numerical clustering” refers to hierarchical and nonhierarchical cluster analysis, latent-class analysis, finite mixture models, and so on.

Part 4 summarizes commercial and public-domain data mining systems and software libraries. Part 5 is devoted to the interdisciplinary links of KDD and includes chapters on these fields. Hand, a prominent statistician, contributes a chapter on data mining’s relationship with statistics. This chapter is particularly useful for scholars educated in more traditional econometric and statistical methods. Part 6 summarizes how data mining is applied to business problems including marketing, fraud detection, risk analysis, production control, and text mining. Part 7 describes how data mining is used in different industry sectors, including banking and finance, telecommunications, engineering, medicine, pharmacology, environmental sciences, and molecular biology. The banking and telecommunications chapters mention some marketing applications. Part 8 gives case studies from industry, public administration, health care, and science. We were impressed by how the same basic methods could be applied to so many different problems.

The book achieves its goal of providing “fast access to the basic concepts” of data mining, but we were disappointed with the coverage of marketing issues. Sections related to marketing were written by consultants or computer scientists, whereas sections covering other subjects were written by leading scholars. For example, the sections about engineering applications are peppered with references to academic conferences, articles, and monographs, whereas those about marketing applications mostly reference trade books (e.g., Berry and Linoff 1997, 2000) and basic marketing textbooks. Relevant marketing literature on relationship marketing, lifetime value, and direct marketing is not incor-

porated or referenced at all. The need for rigorous research on how data mining can solve marketing problems remains. The academic community in marketing research has the opportunity to make this happen.

EDWARD C. MALTHOUSE  
FRANCIS J. MULHERN  
*Northwestern University*

#### REFERENCES

- Berry, Michael and Gordon Linoff (1997), *Data Mining Techniques: For Marketing, Sales, and Customer Support*. New York: John Wiley & Sons.
- and ——— (2000), *Mastering Data Mining: The Art and Science of Customer Relationship Management*. New York: John Wiley & Sons.
- Breiman, Leo (1994), "Comment on 'Neural Networks: A Review from a Statistical Perspective,' by Bing Cheng and D.M. Titterton," *Statistical Science*, 9 (1), 38–42.
- (2001), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16 (3), 199–231.
- Hand, David (1998), "Data Mining: Statistics and More," *The American Statistician*, 52 (2), 112–18.

STATISTICAL ANALYSIS WITH MISSING DATA, 2d ed., Roderick J.A. Little and Donald B. Rubin, Hoboken, NJ: John Wiley & Sons, 2002, 408 pages, \$94.95.

Missing data in marketing surveys and databases is a significant problem. Marketing researchers often deal with missing data by means of listwise deletion (deleting observations entirely if they have incomplete data on any variables used in an analysis), which can reduce the validity of inferences when the data are not missing at random. Many alternative approaches worth considering are discussed in *Statistical Analysis with Missing Data*, written by two people who have contributed a substantial amount to the literature on missing data during the past 30 years.

A key assumption underlying the book (p. 8) is that "missingness indicators hide true values that are meaningful for analysis." Little and Rubin begin with a discussion of missing data patterns, such as haphazard item nonresponse and monotone missing data that arise from attrition in longitudinal studies. They make a critical distinction between patterns of missing data and the mechanisms that lead to missing data, such that data are missing completely at random, missing at random, or not missing at random. The opening chapter concludes with a taxonomy of methods for dealing with missing data: procedures based on completely recorded units (the simple listwise deletion of missing values mentioned previously), weighting procedures, imputation-based procedures, and model-based procedures.

Chapter 2 discusses missing data in experiments. Chapter 3 discusses complete-case (listwise deletion) and available-case analysis, including weighting methods. Weighting procedures modify the weights on individual observations to

adjust for nonresponse, as if the nonresponse were part of the design of the sampling. Chapters 4 and 5 discuss imputation methods. Some imputation-based procedures will probably be familiar to many readers, because they use substitution of existing values in the data set (e.g., the census "hot deck" method), mean values, or values based on regression of observations already in the data set.

Model-based imputation procedures are more complex; they use a defined distribution model for the data and infer values for missing data with a posterior distribution or likelihood function. The model-based procedures take up more than half of the book, Chapters 6–15. Model-based procedures attempt to solve the problem that arises in simpler imputation-based procedures, which tend to fill out the data matrix with values that produce lower residual variances in models and therefore result in biased parameters and statistical tests.

This book provides a thorough, sophisticated analysis of the missing data problem. Little and Rubin have packed the book with solid mathematical treatments of method after method, and researchers with good statistical programming skills will find it possible to program the methods in SAS or other programming languages. Some imputation methods already exist in statistical packages, such as the SAS MI multiple imputation procedure, and the book serves as a good reference for those methods. It will be useful to researchers who desire a thorough treatment of modeling methods in missing data with the intention of investigating their utility and robustness and to modeling specialists who want ideas for different approaches to the missing data problem.

However, this book is not for the average market research practitioner. It is not a how-to book. Readers who want tables laying out recommendations about which methods to use in which situations will find themselves disappointed. More practical-oriented researchers may want to consider Allison's (2001) work, which includes a less technical treatment of missing data methods.

DWAYNE BALL  
*University of Nebraska, Lincoln*

#### REFERENCE

- Allison, Paul D. (2001), *Missing Data*, Sage Series on Quantitative Applications in the Social Sciences, Vol. 136. Thousand Oaks, CA: Sage Publications.

#### BOOKS RECEIVED

- Biemer, Paul P. and Lars E. Lybert (2003), *Introduction to Survey Quality*. Hoboken, NJ: John Wiley & Sons.
- McConnell, Ben and Jackie Huba (2003), *Creating Customer Evangelists: How Loyal Customers Become a Volunteer Sales Force*. Chicago: Dearborn Trade Publishing.
- Rao, J.N.K. (2003), *Small Area Estimation*. Hoboken, NJ: John Wiley & Sons.