Data Mining: Concepts and Techniques

By Jiawei Han and Micheline Kamber

Academic Press, Morgan Kaufmann Publishers, 2001
500 pages, list price $54.95
ISBN 1-55860-489-8

Review by:
Fernando Berzal and Nicolás Marín, University of Granada
Department of Computer Science and AI
fberzal@decsai.ugr.es

Mining information from data: A present-day gold rush. Data Mining is a multidisciplinary field which supports knowledge workers who try to extract information in our “data rich, information poor” environment. Its name stems from the idea of *mining* knowledge from large amounts of data. The tools it provides assist us in the discovery of relevant information through a wide range of data analysis techniques. Any method used to extract patterns from a given data source is considered to be a data mining technique.

Han and Kamber’s book provides more than a good starting point for those interested in this eclectic research field. The book surveys techniques for the main tasks data miners have to perform. Most existing data mining texts emphasize the managerial and marketing aspects involved in the adoption of this technology by modern enterprises. In contrast, Han and Kamber’s textbook focuses on issues such as algorithmic efficiency and scalability from a database perspective.

Basic Concepts for Beginners. The evolution of database technology is an essential prerequisite for understanding the need of knowledge discovery in databases (KDD). This evolution is described in the book to present data mining as a natural stage in the data processing history: we have collected data in the early days of computing, we created database management systems in the seventies, we developed advanced data models in the eighties, and, now, we are left with huge databases which have to be automatically analyzed.

Data mining is a pivotal step in the KDD process: the extraction of interesting patterns from a set of data sources (relational, transactional, object-oriented, spatial, temporal, text, and legacy databases, as well as data warehouses and the World Wide Web). The patterns obtained are used to describe concepts, to analyze associations, to build classification and regression models, to cluster data, to model trends in time-series, and to detect outliers (“data objects that do not comply with the general behavior or model of the data”). Since the patterns which are present in data are not all equally useful, interestingness measures are needed to estimate the relevance of the discovered patterns to guide the mining process. Although the book stresses the importance of interestingness measures and it presents the standard simplicity, certainty, utility, and novelty measures, a more in-depth treatment of alternative interestingness measures would be of interest for data miners but is not given.

From the authors’ point of view, data warehousing and multidimensional databases are introduced as desirable intermediate layers between the original data sources and the On-Line Analytical Mining system the user interacts with. OLAM (also known as OLAP mining) integrates on-line analytical processing with data mining.

In the initial chapters of the book, the reader will find an excellent overview of data warehousing concepts and the proposal of an integrated OLAM architecture, as well as an introduction to DMQL (Data Mining Query Language). Microsoft OLE DB for Data Mining is an alternative to this language and it is briefly described in a separate appendix.

Irrespective of whether data warehouses are used or not, input data must be preprocessed in order to reduce the effect of noise, missing values, and inconsistencies before applying data mining algorithms. Data cleaning, data integration, data transformation, data reduction, discretization and concept hierarchies are enabling techniques which help to prepare the data for the mining process. All these techniques are explained in the book without focusing too much on implementation details so that the reader can easily understand these preprocessing methods.

Data Mining in Action. According to their final goal, data mining techniques can be considered to be descriptive or predictive: Descriptive data mining intends to summarize data and to highlight their interesting properties, while predictive data mining aims to build models to forecast future behaviors.

Generalization is the basis of descriptive techniques and can be used to summarize data by applying attribute-oriented induction using characteristic rules and generalized relations. Analytical characterization is used to perform attribute relevance measurements to identify irrelevant and weakly relevant attributes (the lower the number of attributes, the more efficient the mining process). Generalization techniques can also be extended to discriminate among different classes. The authors refer to these techniques as "class comparison mining". The discussion of descriptive techniques is completed with a brief study of statistical measures (i.e. central tendency and data dispersion measures) and their insightful graphical display.

Association rules are midway between descriptive and predictive data mining (maybe closer to descriptive techniques). They find interesting relationships among large sets of data items and are typically used in market basket analysis. The Apriori family of algorithms is presented as the landmark in association rule mining. Several improvements over the original Apriori algorithm are also described. Han et al.'s FP-Growth (SIGMOD 2000) is thoroughly discussed in the book as an alternative to mine association rules without candidate generation, the common-step in all Apriori-like algorithms. Additional extensions to the basic association rule framework are explored, e.g., iceberg queries and multilevel, multidimensional, constraint-based, and quantitative association rules (which, from our point of view, are artificially categorized into quantitative and distance-based association rules when both of them work with quantitative attributes).

Closer to the predictive arena the book deals with the classical machine learning topics of supervised and unsupervised learning. Several classification and regression techniques are introduced taking into account accuracy, speed, robustness, scalability, and interpretability issues.

Decision trees, Bayesian classifiers, and backpropagation neural networks are presented as outstanding classification techniques. The authors also discuss some classification methods based on concepts from association rule mining. Furthermore, the chapter on classification mentions alternative models based on instance-based learning (e.g., k-NN, CBR…), genetic algorithms, rough and fuzzy sets. We believe that this book section would deserve a more detailed treatment (even a whole volume on its own), which should obviously include an extended version of the study of classifier accuracy found at the end of the chapter.

Regression (called prediction by the authors) appears as an extension of the classical classification models. The former deals with continuous values while the latter is intended to work with discrete categories. Linear regression is clearly explained; multiple, nonlinear, generalized linear, and log-linear regression models are only referenced in the text.

With respect to unsupervised learning (i.e., "learning by observation" rather than learning by examples), cluster analysis is precisely treated in Han and Kamber's book. A general framework for the clustering process is presented pointing out how to compute the dissimilarity between objects taking into account the various types of attributes which can characterize them (binary, nominal, ordinal, interval-based, and ratio-scaled). A taxonomy of clustering methods is proposed including examples for each category: partitioning methods (e.g. k-Means and CLARANS), agglomerative and divisive hierarchical methods (such as BIRCH), density-based methods (like DBSCAN), grid-based methods (such as CLIQUE), and model-based methods (like COBWEB). This categorization of clustering algorithms provides an excellent overview of current clustering techniques, although it can be slightly too dense for people who are new to the field.

The book's survey of data mining tasks and techniques is concluded by the discussion of other relevant problems which are as appealing as the previous ones. For example, outlier analysis has important applications in fraud detection, exception handling, and data preprocessing (i.e., to detect measurement errors); while time-series and sequence mining can be useful to detect trends in market indicators and match similar patterns in genome databases. Unfortunately, these interesting techniques are only briefly described in this book.

Space constraints also limit the discussion of data mining in complex types of data, such as object-oriented databases, spatial, multimedia, and text databases. Web mining, for instance, is only overviewed in its three flavors: web content mining (search engines and information retrieval), web structure mining (linkage analysis), and web usage mining (web log mining).

Practical Issues. The book's final chapter describes some interesting examples of the use of data mining in the real world (i.e., biomedical research, financial data analysis, retail industry, and telecommunication utilities). This chapter also offers some practical tips on how to choose a particular data mining system, advocating for multi-dimensional thinking (as Tom Gilb did in "Principles of Software Engineering Management" some time ago). Moreover, the features of some commercial data mining systems are outlined, such as the authors' DBMiner, whose architecture and capabilities are introduced in a separate appendix. Some buzzwordism about the role of data mining and its social impact can be found in this chapter and forecast of future trends is included at its end, although we feel that the authors' forecast ignores the importance of reusable data mining toolkits and frameworks.

Why to Read This Book. Maybe the authors' goal of covering the whole field of data mining hinders a detailed treatment of some of the topics discussed in the book. Data mining has become an important research area in just a few years and its current breadth makes it impossible to fit into a single volume book. The youth of this field might justify the authors' bias we have found in some specific sections (e.g. they strongly advocate for tightly coupled data mining systems discouraging alternative solutions). Anyway, this book is an indispensable road map for those interested in data mining, both researchers and practitioners.

This book constitutes a superb example of how to write a technical textbook with didactic content and academic rigor. It is written in a direct style with questions and answers scattered throughout the text that keep the reader involved and explain the reasons behind every decision. The presence of examples make concepts easy to understand and the summary and exercises at the end of each chapter support the reader in checking his/her comprehension of the book's contents. The chapters are mostly self-contained, so they can be separately used to teach particular data mining areas. In fact, you may even use the book artwork which is freely available from the Web. Moreover, the bibliographical discussions presented at the end of every chapter describe related work and may prove invaluable for those interested in further reading. A must-have for data miners!