Book Review

Data Mining: Concepts and Techniques by J. Han and M. Kamber. (San Francisco, CA: Morgan Kaufmann, 2001, 550 pp., ISBN 1-55860-489-8). *Reviewed by Zhi-Hua Zhou*.

Written from a *database perspective*, this book is organized into 10 chapters. Chapter 1 provides an introduction to data mining. Chapter 2 focuses on data warehouse and on-line analytical processing. Chapter 3 presents techniques for preprocessing the data prior to mining. Chapter 4 introduces the primitives of data mining that define the specification of a data mining task. Chapter 5 is devoted to descriptive data mining. Chapter 6 deals with association rule mining. Chapter 7 presents techniques for classification and regression. Chapter 8 is on cluster analysis. Chapter 9 focuses on data mining in advanced data repository systems. Chapter 10 discusses applications and challenges of data mining.

What is impressive about this book is that it covers almost all aspects of concepts and techniques of data mining. The bibliography contains more than 400 literatures. Useful bibliographical notes are provided at the end of each chapter, presenting a roadmap for readers who want to learn more from the literature. An extensive index occupying 18 pages, making this book a good reference book or handbook for data mining researchers.

Moreover, this book provides a lot of algorithms geared to the discovery of data patterns hidden in large, real databases. Those algorithms are illustrated in pseudo-code and are easy to be translated into concrete programming languages. This may be especially helpful to data mining practitioners.

Furthermore, this book has good features to serve as a textbook for data mining course. First, the material is presented in a question-and-answer style. Second, each chapter is provided with a set of exercises that could be used as assignments. Third, a suite of slides are provided at the book homepage (www.cs.sfu.ca/~han/dm_book).

However, in spite of its strengths and attractive features, this book also has some drawbacks.

There are much typos or language errors. Although the authors have provided an erratum at the book homepage, still not much has been listed.

Organization of the book seems a bit chaotic. For example, Section 4.4 discusses the architectures of data mining systems, which has little relation to the main topic of Chapter 4, that is, data mining primitives. It could have been better to merge this section into Section 2.6.

There are some ambiguities in the book. In Section 2.2, sales data are depicted as cubes and this is called *data cube representation*. But later, it is said that they are *cuboids*, and data cube is the lattice of cuboids. Such description may cause novices to panic when they try to learn what a data cube is. In Section 4.1.4, *novelty* is described as an objective interestingness measure. But it should be at least mentioned that *novelty* is more often regarded as a subjective interestingness measure. In Chapter 7, *predication* is used parallel to *classification*. But in general, approximating a real-valued function is referred to as *regression* instead of *prediction*, and *predication* encompasses both *classification* and *regression*.

Some claims in this book may be not very adequate. For example, in Section 5.6.1, the authors list several differences between descriptive mining methods and machine learning methods. Some of them, such as the claim that descriptive mining methods do not explicitly store the negative data while machine learning methods do, are not fair.

Adding some material to this book could be useful. In Section 4.1.4, measures of subjective interestingness, such as *actionability* and *unexpectedness*, could be described giving the reader an idea how the subjective measures look like. In Section 7.4.2, the *Laplacian* correction to Naive Bayesian learning should be presented, which is used when training data of some classes is not available. For example, suppose the class attribute *buys* is binary, and the instances are described by two independent binary attributes, i.e. *student* and *credit*. If all the training instances are positive/negative, then the class label of a new instance (*student*, *credit*) should be determined through comparing probabilities:

$$P(buys) = \frac{\#(student \wedge buys)+1}{\#buys+2} \cdot \frac{\#(credit \wedge buys)+1}{\#buys+2} \cdot \frac{\#buys+1}{\#total+2}$$

$$P(\overline{buys}) = \frac{\#(student \wedge \overline{buys})+1}{\#\overline{buys}+2} \cdot \frac{\#(credit \wedge \overline{buys})+1}{\#\overline{buys}+2} \cdot \frac{\#\overline{buys}+1}{\#total+2}$$

where the number of training instances with property X is denoted as $\#X$, the positive and negative values of attribute Y are represented as Y and $\overline{Y}$ respectively, the total number of training instances is denoted as $\#total$.

Overall, this is a good book that could benefit data mining researchers, practitioners, and anyone who wants to learn something about data mining. It is also qualified to be used as a textbook for classes. However, since there is still much room for improvement, a second edition may be necessary before this book becomes an excellent, or even classical, reference in data mining area.

Z.-H. Zhou is with the National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China (e-mail: zhouzh@nju.edu.cn).