

Book Reviews

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 ten 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 to 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 references. 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 occupies 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 pseudocode and are easy to be translated into concrete programming languages. This may be especially helpful to data mining practitioners. Han and Kamber's book has good features to serve as a textbook for a 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 many typos and 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, the 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 quite 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, 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:

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

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

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 $\#\text{total}$.

Overall, this is a good book that could benefit the 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 the data mining area.

Information Dynamics: Foundations and Applications by Gustavo Deco and Bernd Schurmann (New York: Springer-Verlag Inc., 2001, 281 pp., hardbound, ISBN: 0-387-95047-8).

Reviewed by Francesco Camastra

The study of Complex Dynamical Systems is more and more stimulating the interest of neural network community. Deco and Schurmann book covers information processing, a specific branch of the science of Complex Dynamical Systems. This review is organized in two sections: the Overview where the contents of the book are examined and the Conclusion where the most important features are discussed.

I. OVERVIEW

The book is not intended as a standard textbook on dynamical systems. The work presents an interdisciplinary approach to information dynamics based on two distinct tools, nonlinear dynamics and neural networks. In this way the authors intend to establish a new and consistent theoretical framework for the problem of discovering knowledge in dynamical data. The aim of this book "is to present a detailed and unifying formulation of the theory of parametric, nonparametric, and semiparametric statistical structure extraction based on an information-theoretic approach." The first chapter offers a comprehensive survey of the work. The second chapter gives a general overview of dynamical systems and the basic notions and fundamentals involved in time series analysis. Deterministic dynamical systems are introduced with

Manuscript received May 23, 2002.

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

F. Camastra is with the Department of NFM-DISI, University of Genova, 16146 Genova, Italy (email: camastra@disi.unige.it).

Publisher Item Identifier S 1045-9227(02)06493-7.