Journal of Computing and Information Science in Engineering

Book Review

Data Mining: Concepts, Models, Methods, and Algorithms, by Mehmed Kantardzie. IEEE Press, New York, NY, 2003, 360pp., $74.95, ISBN: 0-471-22852-4

REVIEWED BY: ASHOK N. SRIVASTAVA[1]

This book provides an interesting, readable, and comprehensive treatment of the field of data mining for a reader who is not familiar with the concepts, tools, and algorithms. The book provides a nice introduction to the field and discusses standard algorithms and data processing techniques. The emphasis of the book is on data sources that can be transformed into tabular data. Thus, the issues regarding mining unstructured data sets or direct mining of relational databases are not discussed in detail. With few exceptions, the book does not cover many of the more recent developments in this and related fields. Support vector machines, link analysis, and ensemble methods and other aspects of statistical learning, perhaps due to the broad nature of the intended audience. The effective writing and the liberal use of examples provide an interesting guide to this important field. The reference section is robust, covering key publications. As an additional and valuable resource, the book contains a list of data mining software vendors and a compendium of data mining case studies from both the government and industry.

The book is organized according to the data mining process outlined in the first chapter. This process includes the problem definition phase, the data collection phase, the data processing phase, the model building phase, and finally the interpretation or discovery phase of the project. This overview of the data mining process gives the reader an appreciation for the complexity of building data mining models. Care is taken to describe the iterative nature of the data mining process, and there is a good discussion of how the vast generation of data drives the need for data mining technology.

The next chapter discussed the important topics of data preparation and data reduction, emphasizing the subtleties of dealing with different data representations and handling time-dependent data. The third chapter, on the fascinating topic of data reduction techniques, describes a set of methods that can be applied to reduce the dimension of the data set. The author discusses various approaches to dimensionality reduction such as principal components analysis, entropy measures for ranking features, and methods to discretize data.

The fourth and fifth chapters, on learning from data and statistical techniques, provide an overview of the roots of data mining. The chapter on learning from data discusses statistical learning theory at a high level, giving the reader an idea of the issues that arise in finite-sample inductive learning. The author also introduces the idea of supervised and unsupervised learning, and describes the common learning tasks such as clustering, classification, and regression in the framework. The author describes model estimation and methods to control overfitting, i.e., a situation where a model learns the "noise" in the data at the expense of learning the underlying "signal." The fifth chapter describes methods of statistical inference commonly used in data mining applications. It discusses standard methods such as the Naïve Bayes Classifier, linear and logistic regression, and linear discriminant analysis. This chapter would provide a reasonable overview of these large topics for a person unfamiliar with the basic ideas.

The sixth chapter discusses cluster analysis and describes agglomerative hierarchical clustering in some detail, and then discusses the k-means algorithm. These algorithms automatically group data records with similar characteristics together. The direct relationship of the k-means algorithm to Gaussian Mixture Models, which is the underlying probabilistic framework for k-means, is not discussed. The next chapter discusses decision trees for classification problems, specifically focusing on one important algorithm known as C4.5. Decision trees learn a function that maps a data record into one of several predefined classes. The algorithm is discussed in detail, with care being given to the underlying assumptions regarding the classification boundaries.

Following the chapter on decision trees, the text describes association rules, which is an unsupervised learning technique for finding rules that relate the occurrence of one set of items in a data set to the occurrence of another set of items in the same data set. The motivational example that is given is *market-based analysis*, where, for example, a grocer may want to know whether the presence of some items (the collection of those items is called an *item set*) in a consumer's shopping transaction *implies* the presence of other items in the same transaction. This chapter also has a brief discussion of web mining and text mining, to give the reader a flavor for those complex fields.

The ninth chapter provides an introduction to neural networks through their historical roots in brain-style computation. The architecture of the networks are discussed along with methods to optimize both supervised and unsupervised neural network models. Supervised neural networks are generally used for nonlinear regression or classification problems, whereas unsupervised neural networks perform clustering. The tenth chapter discusses genetic algorithms, which are derivative-free methods to perform stochastic optimization when one has a well-defined objective function. The bulk of the chapter is devoted to the description of the class of algorithms, their roots in evolutionary biology, and the implementation of the algorithm. The chapter concludes with a short section on the application of genetic algorithms to a machine learning problem. The eleventh chapter discusses fuzzy logic, and its use in representing uncertainty in data. The chapter concludes with an interesting discussion on extracting fuzzy models from data and describes a process for extracting fuzzy predictive rules from data sets.

The final chapter focuses on visualization techniques, which are a fundamental component of most exploratory data analysis and data mining techniques. The chapter describes the parallel coordi-

[1]Deputy Technical Area Manager, Discovery and System Health Technical Area, Intelligent Systems Division, NASA Ames Research Center.

nates plot and the radial plots, and then moves on to discuss Kohonen maps, which bridge the gap from self-organizing neural networks to visualization techniques.

In summary, *Data Mining: Concepts, Models, Methods, and Algorithms* provides a useful introductory guide to the field of data mining, and covers a broad variety of topics, spanning the space from statistical learning theory, to fuzzy logic, to data visualization. The book is sure to appeal to readers interested in learning about the nuts-and-bolts of the art, science, and practice of data mining.