

BOOK REVIEW

Discovering Knowledge in Data: An Introduction to Data Mining, by D. T. Larose, New York: Wiley, 2005, ISBN 0-471-66657-2, xv + 222 pp., \$69.95.

This book is the first volume of a three-volume series on data mining, which introduces the reader to this rapidly growing field. Data mining, which has gained noticeable popularity in the past decade, is essentially an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization (Cabena et al., 1998) to address the issue of exploring large and complicated databases to identify “interesting” relationships, e.g., high order interactions, or very non-linear relationships that ordinarily would not be detected by standard statistical analyses (Borok, 1997; Szolvits, 1995). This area has been approached by computer scientists and statisticians from slightly different perspectives. The author of the book is a statistician, but has tried to include a computer science theme throughout the book, in which I think he has been successful. As he mentions in the preface, the book is intended to be used either by analysts, managers, and decision makers in industry or as a textbook for an introductory course in data mining for graduate or advanced undergraduate students (in computer science or statistics).

Chapter 1 is a short introductory chapter, in which in addition to a brief description of the Cross-Industry Standard Process for Data Mining (CRISP-DM), several real-world case studies are covered to motivate the topics of subsequent chapters. These case studies are also used to describe the first phase of the CRISP-DM process, namely business understanding.

Chapters 2 and 3 examine the next two phases of the CRISP-DM process, i.e., data understanding and data preparation. Chapter 2 is on data preprocessing, which is divided into the two major tasks of data cleaning and data transformation. In data cleaning, general methods for handling missing data, identifying misclassified records in the data, and also a graphical method for detecting outliers are described. In the data transformation section, min-max normalization and z-score standardization methods are discussed. A numerical method for detecting outliers based on z-score standardization is also covered.

Exploratory data analysis is the topic of Chapter 3, which focuses on data understanding. The chapter begins with making a contrast between hypothesis testing and exploratory data analysis, and is followed by the basic ideas of dealing with correlated variables in the data set. Most of this chapter is dedicated to exploring the variables in a real data set, in which by using several diagrams a number of intuitive approaches for obtaining a high level understanding of the data are proposed.

By using many examples, this exploration covers categorical and numerical variables, multivariate relationships, and selection of interesting subsets of data for further investigation. The idea of binning, along with the basic binning methods, is briefly discussed.

Chapters 4–10 are devoted to the most important phase of the CRISP-DM process, namely the modeling phase. These chapters go over the major data mining tasks in discovering knowledge in data, which is the typical way of approaching the material in most of the existing data mining books (Fayyad et al., 1996; Han and Kamber, 2001; Hastie et al., 2001).

Statistical approaches to estimation and prediction is the subject of Chapter 4, in which some of the more widespread and traditional methods of estimation and prediction are examined. These methods include point estimation and confidence interval estimation. Simple linear regression for dealing with two numerical variables is investigated. Multiple regression is also examined, where the relationship between a response variable and a set of predictor variables is modeled linearly. The provided examples and graphs are chosen carefully, and are of significant help in understanding the material better, especially for those readers who are not already familiar with the discussed methods.

Chapter 5 begins with a brief description of supervised and unsupervised methods in data mining. It is followed by an introduction to classification, which is the most common task in the data mining and machine learning area. Some of the most popular methods for performing classification are explored throughout Chapters 5 to 7. K-nearest neighbor algorithm, which is the simplest, most intuitive, and at the same time one of the most frequently used algorithms for classification is the major focus of Chapter 5. The description of k-nearest neighbor algorithm has brought about the need for an explanation of some basic issues in distance functions which are quite useful and informative.

Chapter 6 is on decision trees. The classification and regression trees (CART) method is introduced as a first algorithm for producing decision trees. As in previous chapters, very detailed examples support the introduced method. This is followed by an introduction to the C4.5 algorithm, a brief comparison with CART, and another well-chosen example. To make the descriptions more concrete, the author uses real-world data to make a comparison between C5.0 (an update of C4.5) and CART.

The introduction to classification methods continues in Chapter 7 where classification based on neural networks is discussed. Gradient descent method, back propagation rules, and sensitivity analysis are discussed in detail. This is intended to help the reader obtain a better understanding of how neural networks work. Although covering other classification methods (e.g., Bayesian classification, support vector machines, and associative classification) would have added value, this flawless chapter on neural networks provides the reader with a good understanding of this class of classification methods.

Hierarchical and k-means clustering are the subject of Chapter 8. Among the large number of existing clustering methods, the author has made the right decision to include the two most useful, but at the same time simple, classes of clustering approaches. The chapter begins with a brief description of measures of similarity and distance between records. Two simple examples of agglomerative hierarchical clustering using nearest-neighbor and farthest-neighbor criteria for determining

the distance between clusters makes this class of algorithms clear for the reader. The description of k-means algorithm is accompanied by illustrative examples as well. The chapter ends with a short section on application of k-means clustering using SAS Enterprise Miner.

Chapter 9 is on Kohonen networks which is another approach used for clustering. Even though performing clustering based on the Kohonen networks is described very well, the level of this chapter seems to be higher than the previous chapters. The author could have devoted the chapter to a more popular and simple clustering algorithm such as DBSCAN (Ester et al., 1996).

Association rule mining, which plays an important role in many business applications including market basket analysis, is discussed in Chapter 10. The description of Apriori, the most well-known association rule mining algorithm, is followed by an information-theoretic approach for extending the discussion to numerical attributes, called generalized rule induction. A few illustrative examples and a clarifying discussion on supervised and unsupervised learning close the chapter.

The last chapter is devoted to techniques for model evaluation, which is an important step in the CRISP process. Model evaluations are of critical importance, since they can help the data miner evaluate the quality and effectiveness of the candidate models before they are actually deployed. This chapter is useful and informative since the author gathers simple evaluation techniques for different data mining tasks such as estimation, prediction, and classification into one chapter.

Adding the names and references to a few of the most popular algorithms for performing each of the discussed data mining tasks (e.g., DBSCAN clustering (Ester et al., 1996), and frequent pattern growth association rule mining (Han et al., 2000)) would have added to the value of the book for the more avid readers. One of the shortcomings of the book is the poor subsection organization which can cause the reader to get lost.

To conclude, the best feature of the book is that, although it is not a comprehensive reference on all the existing and popular data mining methods, the selected subset of methods, is well-chosen. More importantly, all the selected material is described in a simple, clear, and at the same time precise way. Including enough number of real-world case studies, step-by-step examples of real applications, software examples, and screen shots has definitely added to the learning value of the book.

REFERENCES

- Borok, L. S. (1997). Data mining: Sophisticated forms of managed care modeling through artificial intelligence. *J. Health Care Finance* 23:20–36.
- Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A. (1998). *Discovering Data Mining: From Concept to Implementation*. Upper Saddle River, New Jersey: Prentice Hall.
- Ester, M., Kriegel, H., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases. *Proceedings of Knowledge Discovery in Data (KDD'96)*. pp. 226–231.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. Cambridge, Massachusetts: The MIT Press.

- Han, J., Kamber, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco, California: Morgan Kaufmann.
- Han, J, Pei, J., Yin, Y. (2000). Mining frequent patterns without candidate generation. Proceedings of Association for Computing Machinery International Conference on Management of Data (SIGMOD'00). pp. 1–12.
- Hastie, T., Tibshirani R., Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.
- Szolvits, P. (1995). Uncertainty and decisions in medical informatics. *Methods Inform. Med.* 34:111–121.

S. Omid Fatemieh
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801