BOOK REVIEW

by M. Gh. Negoita

"Data Mining Methods for Knowledge Discovery", by K. Cios, W. Pedrycz and R. Swiniarski, Kluwer 1998, 495 pp., http://www.wkap.nl/book.htm/0-7923-8252-8

To quote from the Foreword written by Zdzislaw Pawlak, the "father" of rough sets: "The book probably is the first attempt to provide theoretical basis for knowledge discovery through clear and well-organized way of presenting the fundamentals of several key data mining methods." This excerpt is indisputably an excellent description that underlines the essence of the book. The data mining methods described in the book include an array of fundamental technologies that become profoundly visible in this realm, namely, Rough Sets, Fuzzy Sets, Bayesian Methods, Evolutionary Computing, Machine Learning, Neural networks, Clustering, and Preprocessing. These eight chapters are preceded by Chapter 1 entitled Data Mining and Knowledge Discovery. All but the first chapter include carefully chosen exercises based on the material presented in the chapters. I only wished that the authors provided an accompanying solutions manual. Chapter 1 introduces the field of data mining and knowledge discovery in great detail. It defines a process of knowledge discovery in databases (KDD), its main features, namely, data mining (DM) methods, and the ensuing architectures of KDD. The chapter touches on knowledge representation, basic models and DM, and presents several examples of the KDD systems. The chapter defines the book's domain. It stresses the need to bridge the acute gap existing between present-day data generation capabilities (very fast) and their (poor) comprehension. The authors emphasize a multifaceted character of data mining as an important methodological and algorithmic aspect of this emerging area. This multidisciplinary approach permeating the overall area is the most outstanding feature of data mining: when dealing with the diversity of problems, we need various, quite often complementary approaches. This is what makes machine learning, statistics, pattern recognition etc. different from data mining. This opening chapter also includes, in addition to references, an extensive bibliography of data mining and knowledge discovery. It is an excellent source of information for experts and novices alike.

Next chapter is devoted to rough sets. It is one of the first comprehensive presentations of rough sets theory in the context of classification and feature reduction. The rough sets theory has been developed for knowledge discovery in databases and experimental data sets. This chapter provides an excellent foundation of the rough sets theory with emphasis on revealing and discovering important structures in data, and classification of complex objects. The chapter also provides techniques of feature selection-reduction based on the concept of reduct and core. The rough sets are a powerful foundation for multidimensional pattern processing, feature extraction, reduction, and selection of best features for robust classification and prediction.

Chapter 3 deals with fuzzy sets. Fuzzy sets are linguistic information granules capturing concepts with continuous boundaries. They naturally fit as one among contributing technologies of data mining. Humans operate on information granules rather than on numbers. The size of the information granules implies a certain point of view at data, and help to make associations at this level of specificity (granularity) of information. The chapter covers all pertinent topics that are required to fully understand fuzzy sets and their role in data mining. It starts from the notion of partial membership, defines classes of fuzzy sets and provides fundamental methods aimed at experimental methods of membership estimation. Some basic tools used to manipulate fuzzy sets (such as an extension principle and fuzzy arithmetic) are also covered. A significant portion of the chapter concentrates on the notion of information granularity and its quantification (including topics such as information - based characteristics of fuzzy sets and a frame of cognition). Finally, the chapter contrasts fuzzy sets with rough sets as well as highlights synergistic links between probability and fuzzy sets.

Chapter 4 is on Bayesian methods. It has its roots in probabilistic computing and deals with the statistical Bayesian techniques which occupy an important place in data mining and knowledge discovery. This chapter contains systematic description of Bayesian processing methods for classification. The chapter starts with a two-class two-feature classification, then it provides generalization of methods for multi-feature and multi-class pattern classification. The chapter also discusses classifier design based on discriminant functions for normally distributed probabilities of patterns. Furthermore, the chapter includes major estimation techniques of probability densities used in Bayesian inference. Finally, an important description of probabilistic neural network, as a "hardware" implementation of kernel-based probability density and Bayesian classification, is discussed. This chapter provides also an interesting suite of exercises.

In the sequel (Chapter 5), the authors elaborate on evolutionary computing. This paradigm has gained substantial popularity as a population-based global optimization method, and can be easily found in many areas of application including pattern recognition, control, machine learning, etc. The chapter introduces the underlying concept by looking into the key algorithmic issue of genetic optimization, fitness functions, and various encoding and decoding schemes playing a crucial role in the overall model of evolutionary computing. It also provides a number of illustrative examples as to the form of patterns (especially, rules) and their encoding for the purpose of genetic manipulation. One may wonder how much evolutionary methods are suitable as an optimization vehicle for data mining. On one hand, the computational intensity of evolutionary computing could be easily prohibitive. On the other, the aspect of structural optimization supported therein is highly attractive, especially in a highly heterogeneous environment that begs for a variety of patterns to be explored and eventually discovered in the database. On the whole, once some fundamental dimensionality problems have been tackled, this avenue will rapidly grow in importance as one among the basic tools in the DM toolbox.

Chapter 6 on Machine Learning addresses all major issues of interest in machine learning (ML). It starts with definitions of learning, ways of reducing the number of generated hypotheses, and expands on taxonomy of ML algorithms. Then it provides a section on overfitting which is pertinent not only to ML but to all types of learning algorithms. Then the chapter focuses on inductive learning methods by describing three families of such algorithms: rule algorithms (mainly AQ type), decision tree algorithms (mainly ID type) and their combinations (by presenting the CLIP3 algorithm in detail). The next section describes discretization methods, which is of general interest, not limited to those working in ML. Interesting is the final section on ML in knowledge discovery. It points to problems encountered when working with data bases and very briefly alludes on the ways of overcoming them. The chapter has an extensive appendix, like a case study, describing knowledge discovery form heart images, and results of other DM methods, described elsewhere in the book, on the same data.

Chapter 7 is concerned with neural networks. It describes only few types of neural networks (NN) which, in the opinion of the authors, are most applicable for data mining tasks. After some preliminaries, it starts with a nice description of learning rules and makes a point that the backpropagation learning rule has its origin in stochastic approximation. Next section of the chapter describes in great detail radial basis functions (RBFs). It is one of the best presentations of RBFs I have seen. It covers not only parameter selection methods for the RBFs but also methods how to measure confidence in their outputs. The following sections are devoted to the role of RBFs in knowledge discovery. It describes rule-based interpretation of RBFs, fuzzy context-based RBFs, and NN for feature ranking. All come with graphical illustrations and examples. Then the chapter provides a standard description of Kohonen's self-organizing feature maps network, as an unsupervised clustering algorithm. The last part of the chapter is devoted to a detailed description of an Image Recognition NN Algorithm (IRNN) via series of examples. An interesting feature of the IRNN algorithm is a new similarity measure it uses, which is described in the appendix.

Chapter 8 focuses on clustering. Clustering has been around for many decades as a basic mechanism of unsupervised learning. There are literally hundreds of the clustering algorithms documented in the literature. This chapter discusses the main categories of clustering techniques that seem to be the most pertinent in the setting of data mining. The discussion includes such classes of the methods as hierarchical clustering and objective function-based clustering. The emphasis is on two aspects of clustering that are primordial for data mining, namely the use of some prior knowledge and the minimization of computational overhead associated with finding structures in highly dimensional data. Clustering with partial supervision is one of the solutions to the first problem. The modularization effect delivered by an original context-based clustering is an interesting way of reducing computational effort. There is no doubt that these are two among possible ways of making clustering a key technology of data mining. The chapter points at a right direction and properly identifies the problems yet I should point out that it covers only a peak of the clustering iceberg. The last chapter is devoted to preprocessing. It discusses basics of data preprocessing. First, the chapter describes patterns, data types and then presents examples of preprocessing sequences in knowledge discovery. Next, brief descriptions of basic preprocessing operations are provided. Then the chapter discusses feature selection methods. As an illustration of major preprocessing operations, the chapter presents an excellent description of the Principal Component Analysis (PCA) for pattern projection, feature extraction and reduction, and Fisher's linear discriminant and linear transformation. It also provides a description of the sequence of PCA and Fisher's transformation for feature extraction and reduction. Finally, the chapter presents results of numerical experiments related to texture image classification.

To summarize, the book by Cios, Pedrycz and Swiniarski, is a balanced, thoroughly organized, readable, and highly indispensable source of data mining methods for knowledge discovery. Although being primarily aimed at the senior undergraduate and graduate students, the book can easily please researchers and practitioners working in the area. The authors have underpinned the essence of the most pivotal features of data mining regarded as a highly heterogeneous area of research that exploits a significant number of diverse information technologies in an attempt of making sense of data. This leitmotiv permeates the book as the authors re-examine and take advantage of the well-known approaches and unveil the most important facets of their synergy in the development of data mining algorithms. The book is a pioneering endeavor on the data mining market. An important and highly recommended reading to all seriously involved in data mining for knowledge discovery.

Book reviewed by

Prof.dr. Mircea Gh. Negoita

Wellington, New Zealand