Data Mining

A Knowledge Discovery Approach

Krzysztof J. Cios
Witold Pedrycz
Roman W. Swiniarski
Lukasz A. Kurgan

Springer

Table of Contents

Foreword

“If you torture the data long enough, Nature will confess,” said 1991 Nobel-winning economist Ronald Coase. The statement is still true. However, achieving this lofty goal is not easy. First, “long enough” may, in practice, be “too long” in many applications and thus unacceptable. Second, to get “confession” from large data sets one needs to use state-of-the-art “torturing” tools. Third, Nature is very stubborn — not yielding easily or unwilling to reveal its secrets at all.

Fortunately, while being aware of the above facts, the reader (a data miner) will find several efficient data mining tools described in this excellent book. The book discusses various issues connecting the whole spectrum of approaches, methods, techniques and algorithms falling under the umbrella of data mining. It starts with data understanding and preprocessing, then goes through a set of methods for supervised and unsupervised learning, and concludes with model assessment, data security and privacy issues. It is this specific approach of using the knowledge discovery process that makes this book a rare one indeed, and thus an indispensable addition to many other books on data mining.

To be more precise, this is a book on knowledge discovery from data. As for the data sets, the easy-to-make statement is that there is no part of modern human activity left untouched by both the need and the desire to collect data. The consequence of such a state of affairs is obvious. We are surrounded by, or perhaps even immersed in, an ocean of all kinds of data (such as measurements, images, patterns, sounds, web pages, tunes, etc.) that are generated by various types of sensors, cameras, microphones, pieces of software and/or other human-made devices. Thus we are in dire need of automatically extracting as much information as possible from the data that we more or less wisely generate. We need to conquer the existing and develop new approaches, algorithms and procedures for knowledge discovery from data. This is exactly what the authors, world-leading experts on data mining in all its various disguises, have done. They present the reader with a large spectrum of data mining methods in a gracious and yet rigorous way.

To facilitate the book’s use, I offer the following *roadmap* to help in:

a) reaching certain desired destinations without undesirable wandering, and
b) getting the basic idea of the breadth and depth of the book.

First, an overview: the volume is divided into seven parts (the last one being Appendices covering the basic mathematical concepts of Linear Algebra, Probability Theory, Lines and Planes in Space, and Sets). The main body of the book is as follows: Part 1, Data Mining and Knowledge Discovery Process (two Chapters), Part 2, Data Understanding (three Chapters), Part 3, Data Preprocessing (three Chapters), Part 4, Data Mining: Methods for Constructing Data Models (six Chapters), Part 5, Data Models Assessment (one Chapter), and Part 6, Data Security and Privacy Issues (one Chapter). Both the ordering of the sections and the amount of material devoted to each particular segment tells a lot about the authors’ expertise and perfect control of the data mining field. Namely, unlike many other books that mainly focus on the modeling part, this volume discusses all the important — and elsewhere often neglected — parts before and after modeling. This breadth is one of the great characteristics of the book.

A dive into particular sections of the book unveils that Chapter 1 defines what data mining is about and stresses some of its unique features, while Chapter 2 introduces a Knowledge Discovery Process (KDP) as a process that seeks new knowledge about an application domain. Here, it is pointed out that Data Mining (DM) is just one step in the KDP. This Chapter also reminds us that the KDP consists of multiple steps that are executed in a sequence, where the next step is initiated upon successful completion of the previous one. It also stresses the fact that the KDP stretches between the task of understanding of the project domain and data, through data preparation and analysis, to evaluation, understanding and application of the generated knowledge. KDP is both highly iterative (there are many repetitions triggered by revision processes) and interactive. The main reason for introducing the process is to formalize knowledge discovery (KD) projects within a common framework, and emphasize independence of specific applications, tools, and vendors. Five KDP models are introduced and their strong and weak points are discussed. It is acknowledged that the data preparation step is by far the most time-consuming and important part of the KDP.

Chapter 3, which opens Part 2 of the book, tackles the underlying core subject of the book, namely, data and data sets. This includes an introduction of various data storage techniques and of the issues related to both the quality and quantity of data used for data mining purposes. The most important topics discussed in this Chapter are the different data types (numerical, symbolic, discrete, binary, nominal, ordinal and continuous). As for the organization of the data, they are organized into rectangular tables called data sets, where rows represent objects (samples, examples, patterns) and where columns represent features/attributes, i.e., the input dimension that describes the objects. Furthermore, there are sections on data storage using databases and data warehouses. The specialized data types — including transactional data, spatial data, hypertext, multimedia data, temporal data and the World Wide Web — are not forgotten either. Finally, the problems of scalability while faced with a large quantity of data, as well as the dynamic data and data quality problems (including imprecision, incompleteness, redundancy, missing values and noise) are also discussed. At the end of each and every Chapter, the reader can find good bibliographical notes, pointers to other electronic or written sources, and a list of relevant references.

Chapter 4 sets the stage for the core topics covered in the book, and in particular for Part 4, which deals with algorithms and tools for concepts introduced herein. Basic learning methods are introduced here (unsupervised, semi-supervised, supervised, reinforcement) together with the concepts of classification and regression.

Part 2 of the book ends with Chapter 5, which covers knowledge representation and its most commonly encountered schemes such as rules, graphs, networks, and their generalizations. The fundamental issue of abstraction of information captured by information granulation and resulting information granules is discussed in detail. An extended description is devoted to the concepts of fuzzy sets, granularity of data and granular concepts in general, and various other set representations, including shadow and rough sets. The authors show great care in warning the reader that the choice of a certain formalism in knowledge representation depends upon a number of factors and that while faced with an enormous diversity of data the data miner has to make prudent decisions about the underlying schemes of knowledge representation.

Part 3 of the book is devoted to *data preprocessing* and contains three Chapters. Readers interested in Databases (DB), Data Warehouses (DW) and On-Line Analytical Processing (OLAP) will find all the basics in Chapter 6, wherein the elementary concepts are introduced. The most important topics discussed in this Chapter are Relational DBMS (RDBMS), defined as a collection of interrelated data and a set of software programs to access those data; SQL, described as a declarative language for writing queries for a RDBMS; and three types of languages to retrieve and manipulate data: Data Manipulation Language (DML), Data Definition Language (DDL), and Data Control Language (DCL), which are implemented using SQL. DW is introduced as a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process. Three types of DW are distinguished: virtual data warehouse, data mart, and enterprise warehouse. DW is based on a multidimensional data model: the data is visualized using a multidimensional data cube, in contrast to the relational table that is used in the RDBMS. Finally, OLAP is discussed with great care to details. This Chapter is relatively unique, and thus enriching, among various data mining books that typically skip these topics.

If you are like the author of this Foreword, meaning that you love mathematics, your heart will start beating faster while opening Chapter 7 on *feature extraction (FE) and feature selection (FS) methods*. At this point, you can turn on your computer, and start implementing some of the many models nicely introduced and explained here. The titles of the topics covered reveal the depth and breadth of supervised and unsupervised techniques and approaches presented: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Karhunen-Loeve Transformation, Fisher's linear discriminant, SVD, Vector quantization, Learning vector quantization, Fourier transform, Wavelets, Zernike moments, and several feature selection methods. Because FE and FS methods are so important in data preprocessing, this Chapter is quite extensive.

Chapter 8 deals with one of the most important, and often required, preprocessing methods, the overall goal of which is to reduce the complexity of the data for further data mining tasks. It introduces unsupervised and supervised discretization methods of continuous data attributes. It also outlines a dynamic discretization algorithm and includes a comparison between several state of the art algorithms.

Part 4, *Data Mining: Methods for Constructing Data Models,* is comprised of two Chapters on the basic types of unsupervised learning, namely, Clustering and Association Rules; three Chapters on supervised learning, namely Statistical Methods, Decision Trees and Rule Algorithms, and Neural Networks; and a Chapter on Text Mining. Part 4, along with Parts 3 and 6, forms the core algorithmic section of this great data mining volume. You may switch on your computer again and start implementing various data mining tools clearly explained here.

To show the main features of every Chapter in Part 4, let us start with Chapter 9, which covers clustering, a predominant technique used in unsupervised learning. A spectrum of clustering methods is introduced, elaborating on their conceptual properties, computational aspects and scalability. The treatment of huge databases through mechanisms of sampling and distributed clustering is discussed as well. The latter two approaches are essential for dealing with large data sets.

Chapter 10 introduces the other key unsupervised learning technique, namely, association rules. The topics discussed here are association rule mining, storing of items using transactions, the association rules categorization as single-dimensional and multidimensional, Boolean and quantitative, and single-level and multilevel, their measurement by using support, confidence, and correlation, and the association rules generation from frequent item sets (a priori algorithm and its modifications including: hashing, transaction removal, data set partitioning, sampling, and mining frequent item sets without generation of candidate item sets).

Chapter 11 constitutes a gentle encounter with *statistical methods* for *supervised learning*, which are based on exploitation of probabilistic knowledge about data. This becomes particularly visible in the case of Bayesian methods. The statistical classification schemes exploit concepts of conditional probabilities and prior probabilities — all of which encapsulate knowledge about statistical characteristics of the data. The Bayesian classifiers are shown to be optimal given known probabilistic characteristics of the underlying data. The role of effective estimation procedures is emphasized and estimation techniques are discussed in detail. Chapter 11 introduces regression models too, including both linear and nonlinear regression. Some of the most representative generalized regression models and augmented development schemes are covered in detail.

Chapter 12 continues along statistical lines as it describes main types of inductive machine learning algorithms: decision trees, rule algorithms, and their hybrids. Very detailed description of these topics is given and the reader will be able to implement them easily or come up with their extensions and/or improvements. Comparative performances and discussion of the advantages and disadvantages of the methods on several data sets are also presented here.

The classical statistical approaches end here, and neural network models are presented in Chapter 13. This Chapter starts with presentation of biological neuron models: the spiking neuron model and a simple neuron model. This section leads to presentation of learning/plasticity rules used to update the weights between the interconnected neurons, both in networks utilizing the spiking and simple neuron models. Presentation of the most important neuron models and learning rules are unique characteristics of this Chapter. Popular neural network topologies are reviewed, followed by an introduction of a powerful Radial Basis Function (RBF) neural network that has been shown to be very useful in many data mining applications. Several aspects of the RBF are introduced, including its most important characteristic of being similar (almost practically equivalent) to the system of fuzzy rules.

In Chapter 14, concepts and methods related to text mining and information retrieval are presented. The most important topics discussed are information retrieval (IR) systems that concern an organization and retrieval of information from large collections of semi-structured or unstructured text-based databases and the World Wide Web, and how the IR system can be improved by latent semantic indexing and relevance feedback.

Part 5 of the book consists of Chapter 15, which discusses and explains several important and indispensable model selection and model assessment methods. The methods are divided into four broad categories: data re-use, heuristic, formal, and interestingness measures. The Chapter provides justification for why one should use methods from these different categories on the same data. The Akaike's information criterion and Bayesian information criterion methods are also discussed in order to show their relationship to the other methods covered.

The final part of the book, Part 6, and its sole Chapter 16, treats topics that are not usually found in other data mining books but which are very relevant and deserve to be presented to readers. Specifically, several issues of data privacy and security are raised and cast in the setting of data mining. Distinct ways of addressing them include data sanitation, data distortion, and cryptographic methods. In particular, the focus is on the role of information granularity as a vehicle for carrying out collaborative activities (such as clustering) while not releasing detailed numeric data. At this point, the roadmap is completed.

A few additional remarks are still due. The book comes with two important teaching tools that make it an excellent textbook. First, there is an *Exercises* section at the end of each and every Chapter expanding the volume beyond a great research monograph. The exercises are designed to augment the basic theory presented in each Chapter and help the reader to acquire practical skills and understanding of the algorithms and tools. This organization is suitable for both a textbook in a formal course and for self-study. The second teaching tool is a set of PowerPoint presentations, covering the material presented in all sixteen Chapters of the book.

All of the above makes this book a thoroughly enjoyable and solid read. I am sure that no data miner, scientist, engineer and/or interested layperson can afford to miss it.

Vojislav Kecman
University of Auckland
New Zeland