DATA MINING

Concepts, Models, Methods, and Algorithms

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

Technical Reviewers

Mariofanna Milanova, Professor
Computer Science Department
University of Arkansas at Little Rock
Little Rock, Arkansas, USA

Jozef Zurada, Ph.D.
Professor of Computer Information Systems
College of Business
University of Louisville
Louisville, Kentucky, USA

Witold Pedrycz
Department of ECE
University of Alberta
Edmonton, Alberta, Canada

DATA MINING

Concepts, Models, Methods, and Algorithms

SECOND EDITION

Mehmed Kantardzic
University of Louisville

IEEE

IEEE PRESS

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Kantardzic, Mehmed.
Data mining : concepts, models, methods, and algorithms / Mehmed Kantardzic. – 2nd ed.
p. cm.
ISBN 978-0-470-89045-5 (cloth)
1. Data mining. I. Title.
QA76.9.D343K36 2011
006.3'12–dc22
2011002190

oBook ISBN: 978-1-118-02914-5
ePDF ISBN: 978-1-118-02912-1
ePub ISBN: 978-1-118-02913-8

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To Belma and Nermin

CONTENTS

PREFACE TO THE SECOND EDITION

In the seven years that have passed since the publication of the first edition of this book, the field of data mining has made a good progress both in developing new methodologies and in extending the spectrum of new applications. These changes in data mining motivated me to update my data-mining book with a second edition. Although the core of material in this edition remains the same, the new version of the book attempts to summarize recent developments in our fast-changing field, presenting the state-of-the-art in data mining, both in academic research and in deployment in commercial applications. The most notable changes from the first edition are the addition of

- new topics such as ensemble learning, graph mining, temporal, spatial, distributed, and privacy preserving data mining;
- new algorithms such as Classification and Regression Trees (CART), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Balanced and Iterative Reducing and Clustering Using Hierarchies (BIRCH), PageRank, AdaBoost, support vector machines (SVM), Kohonen self-organizing maps (SOM), and latent semantic indexing (LSI);
- more details on practical aspects and business understanding of a data-mining process, discussing important problems of validation, deployment, data understanding, causality, security, and privacy; and
- some quantitative measures and methods for comparison of data-mining models such as ROC curve, lift chart, ROI chart, McNemar's test, and K-fold cross validation paired t-test.

Keeping in mind the educational aspect of the book, many new exercises have been added. The bibliography and appendices have been updated to include work that has appeared in the last few years, as well as to reflect the change in emphasis when a new topic gained importance.

I would like to thank all my colleagues all over the world who used the first edition of the book for their classes and who sent me support, encouragement, and suggestions to put together this revised version. My sincere thanks are due to all my colleagues and students in the Data Mining Lab and Computer Science Department for their reviews of this edition, and numerous helpful suggestions. Special thanks go to graduate students Brent Wenerstrom, Chamila Walgampaya, and Wael Emara for patience in proofreading this new edition and for useful discussions about the content of new chapters, numerous corrections, and additions. To Dr. Joung Woo Ryu, who helped me enormously in the preparation of the final version of the text and all additional figures and tables, I would like to express my deepest gratitude.

I believe this book can serve as a valuable guide to the field for undergraduate, graduate students, researchers, and practitioners. I hope that the wide range of topics covered will allow readers to appreciate the extent of the impact of data mining on modern business, science, even the entire society.

Mehmed Kantardzic
Louisville
July 2011

PREFACE TO THE FIRST EDITION

The modern technologies of computers, networks, and sensors have made data collection and organization an almost effortless task. However, the captured data need to be converted into information and knowledge from recorded data to become useful. Traditionally, the task of extracting useful information from recorded data has been performed by analysts; however, the increasing volume of data in modern businesses and sciences calls for computer-based methods for this task. As data sets have grown in size and complexity, so there has been an inevitable shift away from direct hands-on data analysis toward indirect, automatic data analysis in which the analyst works via more complex and sophisticated tools. The entire process of applying computer-based methodology, including new techniques for knowledge discovery from data, is often called data mining.

The importance of data mining arises from the fact that the modern world is a data-driven world. We are surrounded by data, numerical and otherwise, which must be analyzed and processed to convert it into information that informs, instructs, answers, or otherwise aids understanding and decision making. In the age of the Internet, intranets, data warehouses, and data marts, the fundamental paradigms of classical data analysis are ripe for changes. Very large collections of data—millions or even hundred of millions of individual records—are now being stored into centralized data warehouses, allowing analysts to make use of powerful data mining methods to examine data more comprehensively. The quantity of such data is huge and growing, the number of sources is effectively unlimited, and the range of areas covered is vast: industrial, commercial, financial, and scientific activities are all generating such data.

The new discipline of data mining has developed especially to extract valuable information from such huge data sets. In recent years there has been an explosive growth of methods for discovering new knowledge from raw data. This is not surprising given the proliferation of low-cost computers (for implementing such methods in software), low-cost sensors, communications, and database technology (for collecting and storing data), and highly computer-literate application experts who can pose "interesting" and "useful" application problems.

Data-mining technology is currently a hot favorite in the hands of decision makers as it can provide valuable hidden business and scientific "intelligence" from large amount of historical data. It should be remembered, however, that fundamentally, data mining is not a new technology. The concept of extracting information and knowledge discovery from recorded data is a well-established concept in scientific and medical studies. What is new is the convergence of several disciplines and corresponding technologies that have created a unique opportunity for data mining in scientific and corporate world.

The origin of this book was a wish to have a single introductory source to which we could direct students, rather than having to direct them to multiple sources. However, it soon became apparent that a wide interest existed, and potential readers other than our students would appreciate a compilation of some of the most important methods, tools, and algorithms in data mining. Such readers include people from a wide variety of backgrounds and positions, who find themselves confronted by the need to make sense of large amount of raw data. This book can be used by a wide range of readers, from students wishing to learn about basic processes and techniques in data mining to analysts and programmers who will be engaged directly in interdisciplinary teams for selected data mining applications. This book reviews state-of-the-art techniques for analyzing enormous quantities of raw data in a high-dimensional data spaces to extract new information useful in decision-making processes. Most of the definitions, classifications, and explanations of the techniques covered in this book are not new, and they are presented in references at the end of the book. One of the author's main goals was to concentrate on a systematic and balanced approach to all phases of a data mining process, and present them with sufficient illustrative examples. We expect that carefully prepared examples should give the reader additional arguments and guidelines in the selection and structuring of techniques and tools for his or her own data mining applications. A better understanding of the implementational details for most of the introduced techniques will help challenge the reader to build his or her own tools or to improve applied methods and techniques.

Teaching in data mining has to have emphasis on the concepts and properties of the applied methods, rather than on the mechanical details of how to apply different data mining tools. Despite all of their attractive "bells and whistles," computer-based tools alone will never provide the entire solution. There will always be the need for the practitioner to make important decisions regarding how the whole process will be designed, and how and which tools will be employed. Obtaining a deeper understanding of the methods and models, how they behave, and why they behave the way they do is a prerequisite for efficient and successful application of data mining technology. The premise of this book is that there are just a handful of important principles and issues in the field of data mining. Any researcher or practitioner in this field needs to be aware of these issues in order to successfully apply a particular methodology, to understand a method's limitations, or to develop new techniques. This book is an attempt to present and discuss such issues and principles and then describe representative and popular methods originating from statistics, machine learning, computer graphics, data bases, information retrieval, neural networks, fuzzy logic, and evolutionary computation.

In this book, we describe how best to prepare environments for performing data mining and discuss approaches that have proven to be critical in revealing important patterns, trends, and models in large data sets. It is our expectation that once a reader has completed this text, he or she will be able to initiate and perform basic activities in all phases of a data mining process successfully and effectively. Although it is easy to focus on the technologies, as you read through the book keep in mind that technology alone does not provide the entire solution. One of our goals in writing this book was to minimize the hype associated with data mining. Rather than making false promises that overstep the bounds of what can reasonably be expected from data mining, we have tried to take a more objective approach. We describe with enough information the processes and algorithms that are necessary to produce reliable and useful results in data mining applications. We do not advocate the use of any particular product or technique over another; the designer of data mining process has to have enough background for selection of appropriate methodologies and software tools.

Mehmed Kantardzic
Louisville
August 2002