Data Mining

Intelligent Systems Reference Library, Volume 12

Editors-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Prof. Lakhmi C. Jain
University of South Australia
Adelaide
Mawson Lakes Campus
South Australia 5095
Australia
E-mail: Lakhmi.jain@unisa.edu.au

Further volumes of this series can be found on our homepage: springer.com

Vol. 1. Christine L. Mumford and Lakhmi C. Jain (Eds.)
Computational Intelligence: Collaboration, Fusion and Emergence, 2009
ISBN 978-3-642-01798-8

Vol. 2. Yuehui Chen and Ajith Abraham
Tree-Structure Based Hybrid Computational Intelligence, 2009
ISBN 978-3-642-04738-1

Vol. 3. Anthony Finn and Steve Scheding
Developments and Challenges for Autonomous Unmanned Vehicles, 2010
ISBN 978-3-642-10703-0

Vol. 4. Lakhmi C. Jain and Chee Peng Lim (Eds.)
Handbook on Decision Making: Techniques and Applications, 2010
ISBN 978-3-642-13638-2

Vol. 5. George A. Anastassiou
Intelligent Mathematics: Computational Analysis, 2010
ISBN 978-3-642-17097-3

Vol. 6. Ludmila Dymowa
Soft Computing in Economics and Finance, 2011
ISBN 978-3-642-17718-7

Vol. 7. Gerasimos G. Rigatos
Modelling and Control for Intelligent Industrial Systems, 2011
ISBN 978-3-642-17874-0

Vol. 8. Edward H.Y. Lim, James N.K. Liu, and Raymond S.T. Lee
Knowledge Seeker – Ontology Modelling for Information Search and Management, 2011
ISBN 978-3-642-17915-0

Vol. 9. Menahem Friedman and Abraham Kandel
Calculus Light, 2011
ISBN 978-3-642-17847-4

Vol. 10. Andreas Tolk and Lakhmi C. Jain
Intelligence-Based Systems Engineering, 2011
ISBN 978-3-642-17930-3

Vol. 11. Samuli Niiranen and Andre Ribeiro (Eds.)
Information Processing and Biological Systems, 2011
ISBN 978-3-642-19620-1

Vol. 12. Florin Gorunescu
Data Mining, 2011
ISBN 978-3-642-19720-8

Florin Gorunescu

Data Mining

Concepts, Models and Techniques

Springer

Prof. Florin Gorunescu
Chair of Mathematics
Biostatistics and Informatics University of
Medicine and Pharmacy of Craiova
Professor associated to the Department of
Computer Science
Faculty of Mathematics and Computer Science
University of Craiova
Romania
E-mail: gorun@umfcv.ro

ISBN 978-3-642-19720-8 e-ISBN 978-3-642-19721-5

DOI 10.1007/978-3-642-19721-5

Intelligent Systems Reference Library ISSN 1868-4394

Library of Congress Control Number: 2011923211

Typeset & *Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To my family

Preface

Data Mining represents a complex of technologies that are rooted in many disciplines: mathematics, statistics, computer science, physics, engineering, biology, etc., and with diverse applications in a large variety of different domains: business, health care, science and engineering, etc. Basically, data mining can be seen as the science of exploring large datasets for extracting implicit, previously unknown and potentially useful information.

My aim in writing this book was to provide a friendly and comprehensive guide for those interested in exploring this vast and fascinating domain. Accordingly, my hope is that after reading this book, the reader will feel the need to deepen each chapter to learn more details.

This book aims to review the main techniques used in data mining, the material presented being supported with various examples, suggestively illustrating each method.

The book is aimed at those wishing to be initiated in data mining and to apply its techniques to practical applications. It is also intended to be used as an introductory text for advanced undergraduate-level or graduate-level courses in computer science, engineering, or other fields. In this regard, the book is intended to be largely self-contained, although it is assumed that the potential reader has a quite good knowledge of mathematics, statistics and computer science.

The book consists of six chapters, organized as follows:

- The first chapter introduces and explains fundamental aspects about data mining used throughout the book. These are related to: what is data mining, why to use data mining, how to mine data? Data mining solvable problems, issues concerning the modeling process and models, main data mining applications, methodology and terminology used in data mining are also discussed.
- Chapter 2 is dedicated to a short review regarding some important issues concerning data: definition of data, types of data, data quality, and types of data attributes.

- Chapter 3 deals with the problem of data analysis. Having in mind that data mining is an analytic process designed to explore large amounts of data in search of consistent and valuable hidden knowledge, the first step consists in an initial data exploration and data preparation. Then, depending on the nature of the problem to be solved, it can involve anything from simple descriptive statistics to regression models, time series, multivariate exploratory techniques, etc. The aim of this chapter is therefore to provide an overview of the main topics concerning exploratory data analysis.
- Chapter 4 presents a short overview concerning the main steps in building and applying classification and decision trees in real-life problems.
- Chapter 5 summarizes some well-known data mining techniques and models, such as: Bayesian and rule-based classifiers, artificial neural networks, k-nearest neighbors, rough sets, clustering algorithms, and genetic algorithms.
- The final chapter discusses the problem of evaluating the performance of different classification (and decision) models.

An extensive bibliography is included, which is intended to provide the reader with useful information covering all the topics approached in this book.

The organization of the book is fairly flexible, the selection of the topics to be approached being determined by the reader himself (herself), although my hope is that the book will be read entirely.

Finally, I wish this book to be considered just as a "compass" helping the interested reader to sail in the rough sea representing the current information vortex.

December 2010 Florin Gorunescu
Craiova

Contents