

Soft Computing for Knowledge Discovery and Data Mining

Soft Computing for Knowledge Discovery and Data Mining

edited by

Oded Maimon

Tel-Aviv University

Israel

and

Lior Rokach

Ben-Gurion University of the Negev

Israel

 Springer

Oded Maimon
Tel Aviv University
Dept. of Industrial Engineering
69978 TEL-AVIV
ISRAEL
maimon@eng.tau.ac.il

Lior Rokach
Ben-Gurion University
Dept. of Information System Engineering
84105 BEER-SHEVA
ISRAEL
liorrk@bgu.ac.il

Library of Congress Control Number: 2007934794

Soft Computing for Knowledge Discovery and Data Mining
Edited by Oded Maimon and Lior Rokach

ISBN 978-0-387-69934-9

e-ISBN 978-0-387-69935-6

Printed on acid-free paper.

© 2008 Springer Science+Business Media, LLC.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2 1

springer.com

To my family
– O.M.

To my wife Ronit, and my two boys, Yarden and Roy
– L.R.

Preface

The information age has made it easy to store large amounts of data. Data mining is a new and exciting field that tries to solve the crisis of information overload by exploring large and complex bodies of data in order to discover useful patterns. It is of extreme importance because it enables modeling and knowledge extraction from abundant data availability. Therefore theoreticians and practitioners are continually seeking techniques to make the process more efficient, cost-effective and accurate. Among the more promising techniques that have emerged in recent years are soft computing methods such as fuzzy sets, artificial neural networks, genetic algorithms. These techniques exploit a tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low cost solutions. This book shows that the soft computing methods extend the envelope of problems that data mining can solve efficiently.

This book presents a comprehensive discussion of the state of the art in data mining along with the main soft computing techniques behind it. In addition to presenting a general theory of data mining, the book provides an in-depth examination of core soft computing algorithms.

To help interested researchers and practitioners who are not familiar with the field, the book starts with a gentle introduction to data mining and knowledge discovery in databases (KDD) and prepares the reader for the next chapters. The rest of the book is organized into four parts. The first three parts are devoted to the principal constituents of soft computing: neural networks, evolutionary algorithms and fuzzy logic. The last part compiles the recent advances in soft computing and data mining.

This book was written to provide investigators in the fields of information systems, engineering, computer science, statistics and management, with a profound source for the role of soft computing in data mining. In addition, social sciences, psychology, medicine, genetics, and other fields that are interested in solving complicated problems can much benefit from this book. The book can also serve as a reference book for graduate / advanced undergraduate level courses in data mining and machine learning. Practitioners among

the readers may be particularly interested in the descriptions of real-world data mining projects performed with soft-computing.

We would like to thank all authors for their valuable contributions. We would like to express our special thanks to Susan Lagerstrom-Fife and Sharon Palleschi of Springer for working closely with us during the production of this book.

Tel-Aviv, Israel
Beer-Sheva, Israel

Oded Maimon
Lior Rokach

July 2007

Contents

Introduction to Soft Computing for Knowledge Discovery and Data Mining <i>Oded Maimon, Lior Rokach</i>	1
--	---

Part I Neural Network Methods

Neural Networks For Data Mining <i>G. Peter Zhang</i>	17
---	----

Improved SOM Labeling Methodology for Data Mining Applications <i>Arnulfo Azcarraga, Ming-Huei Hsieh, Shan-Ling Pan, Rudy Setiono</i>	45
--	----

Part II Evolutionary Methods

A Review of Evolutionary Algorithms for Data Mining <i>Alex A. Freitas</i>	79
--	----

Genetic Clustering for Data Mining <i>Murilo Coelho Naldi André Carlos Ponce de Leon Ferreira de Carvalho Ricardo José Gabrielli Barreto Campello Eduardo Raul Hruschka</i>	113
---	-----

Discovering New Rule Induction Algorithms with Grammar-based Genetic Programming <i>Gisele L. Pappa, Alex A. Freitas</i>	133
--	-----

Evolutionary Design of Code-matrices for Multiclass Problems <i>Ana Carolina Lorena, André C. P. L. F. de Carvalho</i>	153
--	-----

Part III Fuzzy Logic Methods

The Role of Fuzzy Sets in Data Mining
Lior Rokach 187

Support Vector Machines and Fuzzy Systems
Yixin Chen 205

KDD in Marketing with Genetic Fuzzy Systems
Jorge Casillas, Francisco J. Martínez-López 225

Knowledge Discovery in a Framework for Modelling with Words
Zengchang Qin, Jonathan Lawry 241

Part IV Advanced Soft Computing Methods and Areas

Swarm Intelligence Algorithms for Data Clustering
Ajith Abraham, Swagatam Das, Sandip Roy 279

A Diffusion Framework for Dimensionality Reduction
Alon Schclar 315

Data Mining and Agent Technology: a fruitful symbiosis
Christos Dimou, Andreas L. Symeonidis, Pericles A. Mitkas 327

Approximate Frequent Itemset Mining In the Presence of Random Noise
Hong Cheng, Philip S. Yu, Jiawei Han 363

The Impact of Overfitting and Overgeneralization on the Classification Accuracy in Data Mining
Huy Nguyen Anh Pham, Evangelos Triantaphyllou 391

Index 433

List of Contributors

Ajith Abraham

Center of Excellence for Quantifiable
Quality of Service (Q2S),
Norwegian University of Science and
Technology,
Trondheim, Norway
ajith.abraham@ieee.org

Arnulfo Azcarraga

College of Computer Studies,
De La Salle University, Manila,
The Philippines
azcarraga
@canlubang.dlsu.edu.ph

Ricardo José Gabrielli Barreto Campello

Instituto de Ciências Matemáticas e
de Computação,
Universidade de São Paulo
campello@icmc.usp.br

André Carlos Ponce de Leon Ferreira de Carvalho

Instituto de Ciê
ncias Matemá
ticas e de Computação
Universidade de São Paulo
andre@icmc.usp.br

Jorge Casillas

Dept. of Computer Science and
Artificial Intelligence,
University of Granada,
Spain
casillas@decsai.ugr.es

Yixin Chen

Dept. of Computer and Information
Science
The University of Mississippi
MS 38655
ychen@cs.olemiss.edu

Hong Cheng

University of Illinois at Urbana-
Champaign
hcheng3@cs.uiuc.edu

Swagatam Das

Dept. of Electronics and Telecommu-
nication Engineering,
Jadavpur University,
Kolkata 700032,
India.

Christos Dimou

Electrical and Computer Engineering
Dept.
Aristotle University of Thessaloniki,
54 124, Thessaloniki,
Greece
cdimou@issel.ee.auth.gr

Alex A. Freitas

Computing Laboratory,
University of Kent,
Canterbury, Kent, CT2 7NF, UK
A.A.Freitas@kent.ac.uk

Jiawei Han

University of Illinois at Urbana-
Champaign
hanj@cs.uiuc.edu

Eduardo Raul Hruschka

eduardo.hruschka
@pesquisador.cnpq.br

Ming-Huei Hsieh

Dept. of International Business,
National Taiwan University,
Taiwan
mhhsieh@management.ntu.edu.tw

Jonathan Lawry

Artificial Intelligence Group,
Department of Engineering Mathe-
matics,
University of Bristol,
BS8 1TR, UK.
j.lawry@bris.ac.uk

Ana Carolina Lorena

Centro de Matemática,
Computação e Cognição
Universidade Federal do ABC
Rua Catequese, 242,
Santo André, SP, Brazil
ana.lorena@ufabc.edu.br

Oded Maimon

Dept. of Industrial Engineering
Tel-Aviv University
Israel
maimon@eng.tau.ac.il

Francisco J. Martínez-López

Dept. of Marketing, University of
Granada, Spain
fjmlopez@ugr.es

Murilo Coelho Naldi

Instituto de Ciê
ncias Matemá
ticas e de Computação
Universidade de São Paulo
murilocn@icmc.usp.br

Shan-Ling Pan

School of Computing,
National University of Singapore,
Singapore
pansl@comp.nus.edu.sg

Gisele L. Pappa

Computing Laboratory
University of Kent
Canterbury, Kent, CT2 7NF, UK
glp6@kent.ac.uk

Huy Nguyen Anh Pham

Dept. of Computer Science,
298 Coates Hall,
Louisiana State University,
Baton Rouge, LA 70803
hpham15@lsu.edu

Zengchang Qin

Berkeley Initiative in Soft Comput-
ing (BISC),
Computer Science Division,
EECS Department,
University of California,
Berkeley, CA 94720, US.

`zqin@eecs.berkeley.edu`

Lior Rokach

Dept. of Information System Engineering,
Ben-Gurion University,
Israel
`liorrk@bgu.ac.il`

Sandip Roy

Dept. of Computer Science and Engineering,
Asansol Engineering College,
Asansol-713304, India.

Alon Schclar

School of Computer Science,
Tel Aviv University,
Tel Aviv 69978,
Israel
`shekler@post.tau.ac.il`

Rudy Setiono

School of Computing,
National University of Singapore,
Singapore
`rudys@comp.nus.edu.sg`

Andreas L. Symeonidis

Electrical and Computer Engineering

Dept.

Aristotle University of Thessaloniki,
54 124, Thessaloniki,
Greece
`asymeon@iti.gr`

Pericles A. Mitkas

Electrical and Computer Engineering
Dept.
Aristotle University of Thessaloniki,
54 124, Thessaloniki,
Greece
`mitkas@eng.auth.gr`

Evangelos Triantaphyllou

Dept. of Computer Science,
298 Coates Hall,
Louisiana State University,
Baton Rouge, LA 70803
`trianta@lsu.edu`

Philip S. Yu

IBM T. J. Watson Research Center
`psyu@us.ibm.com`

G. Peter Zhang

Georgia State University,
Dept. of Managerial Sciences
`gpzhang@gsu.edu`