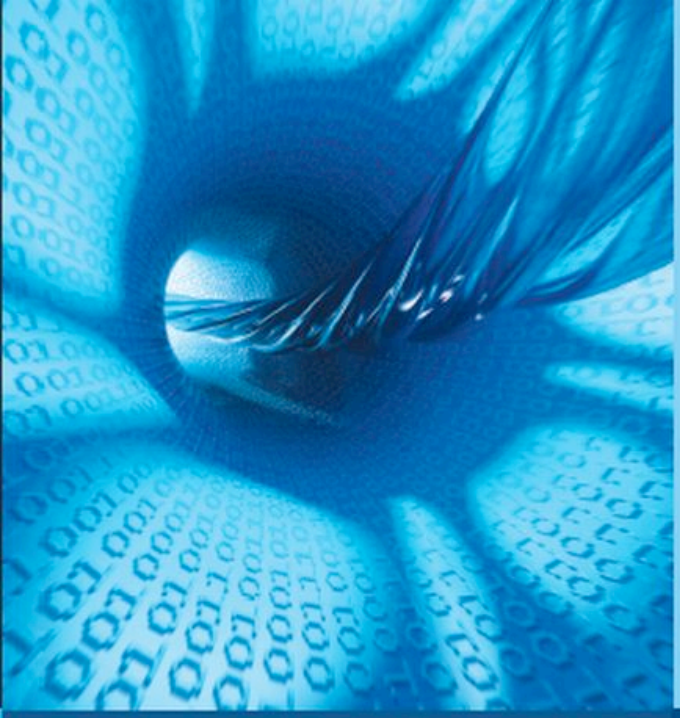

Oded Maimon
Lior Rokach
Editors

Soft Computing for Knowledge Discovery and Data Mining

Springer

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.

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 extreme importance because it enables modeling and knowledge extraction from abundance data availability. Therefore theoreticians and practitioners are continually seeking techniques to make the process more efficient, cost-effective and accurate. Among the more promising technique 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 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 *Oded Maimon*
Beer-Sheva, Israel *Lior Rokach*

July 2007

Contents

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
azcarragaa
@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 Telecommunication 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 Mathematics,
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 Computing (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

Introduction to Soft Computing for Knowledge Discovery and Data Mining

Oded Maimon[1] and Lior Rokach[2]

[1] Department of Industrial Engineering, Tel-Aviv University, Ramat-Aviv 69978, Israel,
maimon@eng.tau.ac.il
[2] Department of Information System Engineering, Ben-Gurion University, Beer-Sheba, Israel,
liorrk@bgu.ac.il

Summary. In this chapter we introduce the Soft Computing areas for Data Mining and the Knowledge Discovery Process, discuss the need for plurality of methods, and present the book organization and abstracts.

1 Introduction

Data Mining is the science, art and technology of exploring data in order to discover insightful unknown patterns. It is a part of the overall process of Knowledge Discovery in Databases (KDD). The accessibility and abundance of information today makes data mining a matter of considerable importance and necessity.

Soft computing is a collection of new techniques in artificial intelligence, which exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low solution cost. Given the history and recent growth of the field, it is not surprising that several mature soft computing methods are now available to the practitioner, including: fuzzy logic, artificial neural networks, genetic algorithms, and swarm intelligence. The aims of this book are to present and explain the important role of soft computing methods in data mining and knowledge discovery.

The unique contributions of this book is in the introduction of soft computing as a viable approach for data mining theory and practice, the detailed descriptions of novel soft-computing approaches in data mining, and the illustrations of various applications solved in soft computing techniques, including: Manufacturing, Medical, Banking, Insurance, Business Intelligence and others. The book does not include some of the most standard techniques in Data Mining, such as Decision Trees (the reader is welcome to our new book, from 2007, dedicated entirely to Decision Trees). The book include the leading soft

computing methods, though for volume reasons it could not cover all methods, and there are further emerging techniques, such as fractal based data mining (a topic of our current research).

Since the information age, the accumulation of data has become easier and storing it inexpensive. It has been estimated that the amount of stored information doubles less than twenty months. Unfortunately, as the amount of electronically stored information increases, the ability to understand and make use of it does not keep pace with its growth. Data Mining is a term coined to describe the process of sifting through large databases for interesting patterns and relationships. The studies today aim at evidence-based modeling and analysis, as is the leading practice in medicine, finance, intelligence and many other fields. Evidently, in the presence of the vast techniques' repertoire and the complexity and diversity of the explored domains, one real challenge today in the data mining field is to know how to utilize this repertoire in order to achieve the best results. The book shows that the soft computing methods extend the envelope of problems that data mining can solve efficiently. The techniques of soft computing are important for researchers in the fields of data mining, machine learning, databases and information systems, engineering, computer science and statistics.

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

The material of this book has been taught by the authors in graduate and undergraduate courses at Tel-Aviv University and Ben-Gurion University. The book can also serve as a reference book for graduate and advanced undergraduate level courses in data mining and machine learning.

In this introductory chapter we briefly present the framework and overall knowledge discovery process in the next two sections, and then the logic and organization of this book, with brief description of each chapter.

2 The Knowledge Discovery process

This book is about methods, which are the core of the Knowledge Discovery process. For completion we briefly present here the process steps. The knowledge discovery process is iterative and interactive, consisting of nine steps.

Note that the process is iterative at each step, meaning that moving back to previous steps may be required. The process has many "artistic" aspects in the sense that one cannot present one formula or make a complete taxonomy for the right choices for each step and application type. Thus it is required to understand the process and the different needs and possibilities in each step.

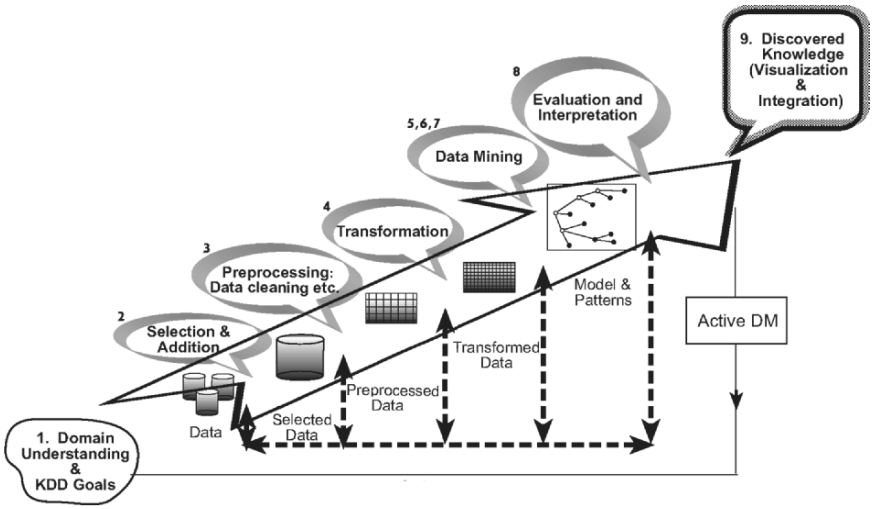

Fig. 1. The Process of Knowledge Discovery in Databases.

The process starts with determining the KDD goals, and "ends" with the implementation of the discovered knowledge. Then the loop is closed - the Active Data Mining part starts. As a result, changes can be made in the application domain (such as offering different features to mobile phone users in order to reduce churning). This closes the loop, and the effects are then measured on the new data repositories, and the KDD process is launched again.

Following is a brief description of the nine-step KDD process, starting with a managerial step:

1. Developing an understanding of the application domain: This is the initial preparatory step. It prepares the scene for understanding what should be done with the many decisions (about transformations, algorithms, representation, etc.). The people who are in charge of a KDD project need to understand and define the goals of the end-user and the environment in which the knowledge discovery process will take place (including relevant prior knowledge). As the KDD process proceeds, there may be even a revision of this step.
   Having understood the KDD goals, the preprocessing of the data starts, defined in the next three steps.
2. Selecting and creating a data set on which discovery will be performed: Having defined the goals, the data that will be used for the knowledge discovery should be determined. This includes finding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one data set, including the attributes