DATA MINING

Methods for

KNOWLEDGE DISCOVERY

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

DATA MINING

Methods for

KNOWLEDGE DISCOVERY

by

Krzysztof J. Cios
University of Toledo
Toledo, OH, USA

Witold Pedrycz
University of Manitoba
Winnipeg, CANADA

Roman W. Swiniarski
San Diego State University
San Diego, CA USA

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

Library of Congress Cataloging-in-Publication Data

Cios, Krzysztof J.
Data mining methods for knowledge discovery / by Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski.
p. cm. -- (Kluwer international series in engineering and computer science ; SECS 458)
Includes bibliographical references and index.
ISBN 978-1-4613-7557-9 ISBN 978-1-4615-5589-6 (eBook)
DOI 10.1007/978-1-4615-5589-6
1. Database management 2. Data mining. I. Pedrycz, Witold, 1953- . II. Świniarski, Roman. III. Title. IV. Series.
QA76.9.D3.C495 1998
006.3--dc21 98-29384
CIP

Originally published by Kluwer Academic Publishers in 1998
Softcover reprint of the hardcover 1st edition 1998
Third Printing 2000.

Printed on acid-free paper.

TO OUR FAMILIES

Authors

Foreword

It is my great pleasure and privilege to welcome the initiative of three prominent researchers to present an excellent book on new, very important and rapidly developing areas of research and applications -- data mining methods for knowledge discovery in databases.

Current indications show that the knowledge discovery processes, and data mining methods they use, undoubtedly belong to the fastest growing domains, and, as such, attract researchers and practitioners world-wide. Data mining methods and the contexts in which they are used for knowledge discovery are relatively new domains, although strongly related to more established areas of machine intelligence and machine learning. However, it is worth noting that their origin can be traced back to the ideas of Bertrand Russell and Karl Popper.

The book presents almost the entire range of data mining methods such as rough sets, fuzzy sets, Bayesian methods, evolutionary computing, machine learning, neural networks, clustering and preprocessing. All are fundamental tools for knowledge discovery in databases. The quest for presenting basic methods for knowledge discovery in databases seems to be of great importance for all interested in making sense of large amounts of data collected almost daily. The book probably is the first attempt to provide a theoretical basis for knowledge discovery through clear and well-organized ways of presenting the fundamentals of several key data mining methods. Hence, the book can serve as a valuable introduction to data mining methods, and as a guide of how to use them for discovering new knowledge. I am confident that the book has a good chance of becoming one of the most referenced books in the area.

I am deeply convinced that the book will play an important role in pursuing further development and applications of the described in the book data mining methods for knowledge discovery -- a truly fascinating research endeavor. The Authors need to be congratulated for their pioneering work.

Zdzislaw Pawlak

PREFACE

We live in an information age. Information has become a very important commodity. Every second hundreds of thousands of new records of information are generated. This information needs to be summarized and synthesized in order to support effective decision-making. In short, there is an urgent need to make sense of large amounts of data.

Data Mining (DM) methods are used in a process called Knowledge Discovery (KD) to reveal new pieces of knowledge from large databases. The terms *knowledge discovery* and *data mining* first appeared in late eightieths and have been used ever since. DM methods and their use for knowledge discovery are the topic of this book. We show how they can be used within the knowledge discovery process and elaborate on interactions between DM methods.

We need to keep in mind that neither of the described data mining methods is a panacea for solving problems involving hundreds of thousands of highly dimensional records. A DM method can work well in some domains but fail in others. This was in fact one of the main reasons for coming up with the new "umbrella" term of data mining and knowledge discovery in databases encompassing databases, machine learning, statistics, Artificial Intelligence, visualization, high performance computing, to .name just a few research areas. It was simply a realization that no single method can be expected to work well with diverse types of large databases. A new methodology was needed.

Let us use an analogy from the business world. In the modern auto factory the next process, in a series of processes required to make a car, is treated as a customer for the current process. Thus, it is paid the highest attention and so is each process in turn. In other words, any improvement made by the current process contributes to the final outcome. The ultimate customer is obviously the buyer of a car. If the final product is not satisfactory, however, the entire chain of processes is altered. This may involve changing the way in which the individual processes operate, or removing some, or adding new ones, or merging several old processes into one. We do essentially the same in a knowledge discovery process where several data mining tools may be used on a database before an interesting piece of new information, or knowledge, is generated. This result is ultimately judged by a decision-maker who

created the database and collected information with some goal in mind. If the new knowledge is found not to be useful we may have to go back and redo what we have already done by using particular DM methods. This can be done either by changing the way in which method were used (parameters, different samples, etc.), or by using other methods, or eliminating those which gave results that could not be successfully used by another method following use of a particular DM method.

Why the interest in data mining methods? There are numerous reasons behind researching fundamental concepts, pursuing algorithmic aspects, and deploying concrete data mining and knowledge discovery systems. In general, the objective of data mining and knowledge discovery is to make sense of data. Depending on the problem at hand, this translates into more specific goal, or set of goals. In business, data captures information about the markets, customers, arising competition, etc. In a manufacturing sector, making sense of data becomes a key to new and successful technologies, higher quality of products, and improved performance of processes. In science, it helps to understand phenomena under study, help establish new directions, or to suggest exploration along new promising paths.

Writing a book on data mining and knowledge discovery is not an easy task. The difficulties can be attributed to many factors. First, the area itself is quite new and not fully defined. New ideas need to emerge regarding how to approach the task of generating new knowledge that goes beyond each research of contributing areas. They need to be evaluated, and put in a meaningful context. As with any new research endeavor, there is some hype and perhaps excessively high expectations. Thus, presentation of the overall material in a balanced way is a difficult task. Thirdly, data mining and knowledge discovery is not a coherent field. It dwells upon many already well established technologies including data cleaning, data preprocessing, machine learning (ML), pattern recognition, statistics, neural networks (NN), fuzzy sets (FS), rough sets (RS), clustering etc. The book strives to achieve a sensible balance between them, show ways in which they interact and construct a homogeneous framework of how they can be used within the knowledge discovery process.

It is not possible to cover the entire scope of data mining methods that can be used at different stages of a discovery process. We had to make some well-thought selections. Our objective was to concentrate on DM methods themselves and discuss available methodologies in to be used within the knowledge discovery process. The available methods are also revisited in terms of their computational efficiency and suitability as the basic tools for knowledge discovery. The organization of the material in the book is shown in Figure 1. It is intended to help the reader select the most suitable methods for solving specific problems.

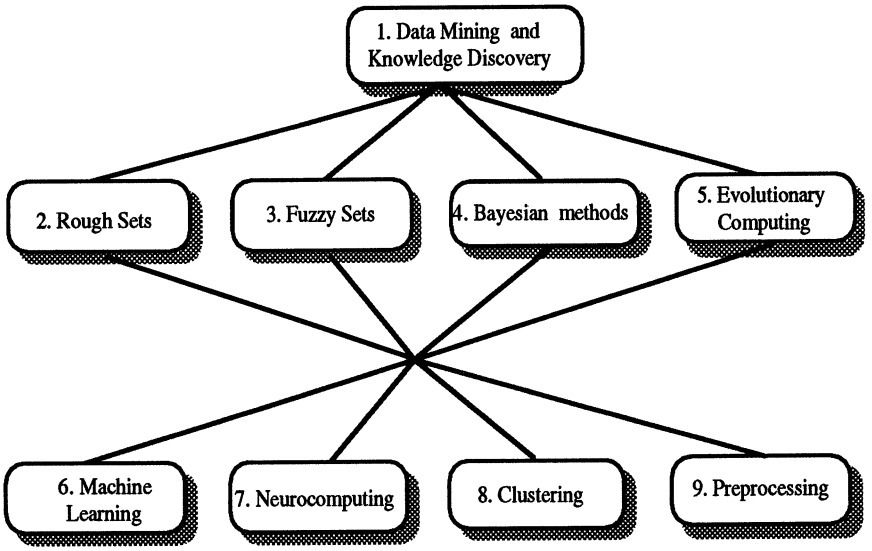

Figure 1. Data mining methods for knowledge discovery: a general roadmap

The material is presented as follows. Chapter 1 is an introduction to data mining and knowledge discovery. It defines the area, elaborates on methodological aspects and summarizes the main groups of the ensuing algorithms. The four chapters shown at the second level in Figure 1 are more basic in nature. Chapters 2 and 3 are set-oriented generalizations aimed at representing and processing uncertainty. They naturally fall under the umbrella of granular computing that becomes an indispensable vehicle of data summarization. An effect of information granularity is achieved by using theories of sets, fuzzy sets, and rough sets. Information granulation can be likened to filtering of the original data through sieves with holes of different sizes. Chapter 4 has its roots in probabilistic computing and deals with the Bayesian techniques which occupy an important place in the knowledge discovery process. Evolutionary techniques presented in Chapter 5 are discussed at the algorithmic level with an intent of portraying them as an essential tool for structural and parametric optimization.

On the third level, as depicted in Figure 1, there are again four chapters. Chapter 6 deals with machine learning (ML) whose role in a knowledge discovery process is indisputable. We present the main ideas of ML with emphasis on use of these techniques in the knowledge discovery process. The chapter also includes a section on discretization of variables. Neural networks have been found to be useful because of their learning and generalization abilities. Chapter 7 discusses them as a part of the overall repertoire of heterogeneous DM technologies. Similarly, the ideas of unsupervised learning (clustering) along with many DM-oriented enhancements are covered in Chapter 8. Essential for the knowledge discovery process as well as for

individual data mining methods is data preprocessing which is described in Chapter 9.

We should also stress here what the book is not about. It is not about databases, data warehousing, and high performance computing. It also does not cover visualization of data and knowledge which help improve man-machine interaction and may lead to interactive knowledge discovery.

The book is aimed at a number of groups of readers: First, it is geared to a broad audience of graduate and undergraduate students. The material is tailored to a one-semester or two-quarter courses on data mining and knowledge discovery. The instructor may like to spend more time on selected approaches, depending on the needs of the students and the research focus of the course. Similarly, if the students have enough prerequisite knowledge about some of the contributing technologies, the corresponding portion of the material could serve as a useful refresher. The book can be also used in senior undergraduate courses on information processing. Second, the material could be of genuine interest to a research community including those interested in entering the area of data mining and knowledge discovery as well as to those who want to systematize and acquire overall picture of the area. Thirdly, the book can also serve as a comprehensive reference. In this sense, it can be used in a number of courses offered to industry and public sector interested in the knowledge discovery process.

The book presents the material in a comprehensive fashion. The material is self-contained. We anticipate that readers have a limited formal mathematical background. Our intent is to make the presentation fully operational so that the algorithms described lend themselves to easy implementation and experimentation. This does not necessarily mean that all details are included; some of them are left out on purpose as they blur the picture. Each chapter comes with a number of exercises that help readers digest the material and extended bibliography. The exercises are at different levels of difficulty. Some are more demanding and could be helpful in inspiring further research activities.

The area of data mining and knowledge discovery is developing at a high pace. To keep up with it, the reader is encouraged to go to cyberspace where there is a vast number of useful information. Some of the Web cites are listed below:

CWI (Centrum voor Wiskunde Informatica),
http://www.cwi.nl/cwi/projects/datamining.html

German National Research Center for Information Technology,
http://orgwis.gmd.de/explora/pages.html

GTE Laboratories,
http://info.gte.com/~kdd and http://info.gte.com/~kdd/kdd-at-gte.html

IBM Almaden, Data Mining project,
http://www.almaden.ibm.com/cs/quest

Information Technology Institute,
http://www.iti.gov.sg//RnD/ia/ia_page.html

Los Alamos National Laboratory,
http://www.acl.lanl.gov/sunrise/DataMining/intro.html

University of Birmingham, United Kingdom,
http://www.cs.bham.ac.uk/~anp/TheDataMine.html

University of Ulster at Jordanstown, http://iserve1.infj.ulst.ac.uk:8080/main.html

Vanderbilt University,
http://www.vuse.vanderbilt.edu/~biswas/ResearchPages/kdd.html

We gratefully acknowledge support from NASA Lewis Research Center and Natural Sciences and Engineering Research Council. Our sincere gratitude goes to Professor Andrzej Skowron for helpful comments and suggestions. We also acknowledge input from our graduate students, in particular Jeff Lovelace, Dorel Sala, and Ruibing Zhang. Gratitude is extended to Scott Delman from Kluwer Academic Publishers for his professional advice and continuous encouragement.

Authors