
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

Copyright © 1998 by Springer Science+Business Media New York

Originally published by Kluwer Academic Publishers in 1998

Softcover reprint of the hardcover 1st edition 1998

Third Printing 2000.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, Springer Science+Business Media, LLC.

Printed on acid-free paper.

TO OUR FAMILIES

Authors

Foreword	xv
----------	----

PREFACE	xvii
---------	------

CHAPTER 1 DATA MINING AND KNOWLEDGE DISCOVERY	1
1.1 Data Mining and Information Age: Emerging Quests	1
1.2 Defining Knowledge Discovery	2
1.3 Architectures of Knowledge Discovery	4
1.4 Knowledge Representation	8
Types of Data	8
Cognitive Perspective and Information Granulation	10
1.5 Main Types of Revealed Patterns	11
1.6 Basic Models of Data Mining	14
1.7 Knowledge Discovery and Related Research Areas	15
1.8 Main Features of a Knowledge Discovery Process	16
1.9 Coping with Reality. Sampling in Databases	17
1.10 Selected Examples of Knowledge Discovery Systems	18
1.11 Summary	20
References	21
Additional Readings	22
 CHAPTER 2 ROUGH SETS	 27
2.1 Introduction	27
2.2 Information System	28
2.3 Indiscernibility Relation	32
2.4 Discernibility Matrix	35
2.5 Decision Tables	37
2.6 Approximation of Sets. Approximation Space	39
Definable and Non-Definable Sets	42
2.7 Accuracy of Approximation	45
Uncertainty and the Rough Membership Function of a Set	46
2.8 Approximation and Accuracy of Classification	47
2.9 Classification and Reduction	49
Reduct and Core	52
2.10 Decision Rules	56
2.11 Dynamic Reducts	60
Decision Rules Computed from Dynamic Reducts	63
2.12 Summary	63
2.13 Exercises	64
References	66
Appendix A2: Algorithms for Finding Minimal Subsets	69
Introduction	69
Discernibility Matrix and Function	70
Reducts in a Discernibility Matrix	70

CHAPTER 3 FUZZY SETS	73
3.1 Introduction	73
3.2 Basic Definition	75
3.3 Types of Membership Functions	77
Triangular Membership Functions	77
S-Membership Function	78
Trapezoidal Membership Function	78
Gaussian Membership Function	78
3.4 Characteristics of a Fuzzy Set	78
3.5 Membership Function Determination	82
Horizontal Method of Membership Estimation	83
Vertical Method of Membership Estimation	84
Pairwise Comparison Method of Membership Function Estimation	84
Problem Specification-Based Membership Determination	86
Membership Estimation as a Problem of Parametric Optimization	87
3.6 Fuzzy Relations	88
3.7 Set Theory Operations and Their Properties	89
Triangular Norms	91
Triangular Norms as the Models of Operations on Fuzzy Sets	95
3.8 The Extension Principle and Fuzzy Arithmetic	97
3.9 Information—Based Characteristics of Fuzzy Sets	98
Entropy Measure of Fuzziness	98
Energy Measure of Fuzziness	102
Specificity of a Fuzzy Set	103
Matching Fuzzy Sets—Possibility and Necessity Measures	104
3.10 Numerical Representation of Fuzzy Sets	105
3.11 Rough Sets and Fuzzy Sets	108
Processing of Fuzzy Sets in the Setting of the Indiscernibility Relation	108
Fuzzy Sets as Elements of the Indiscernibility Relation	108
3.12 The Frame of Cognition	109
Basic Definition	109
Main Properties	110
Information Granularity and the Development of a Frame of Cognition	113
Approximation Aspects of the Frame of Cognition	116
Robustness Properties of the Frame of Cognition	118
3.13 Probability and Fuzzy Sets	121
Hybrid Fuzzy-Probabilistic Models of Uncertainty	122
3.14 Summary	125
3.15 Exercises	125
References	127
 CHAPTER 4 BAYESIAN METHODS	 131
4.1 Introduction	131
4.2 Basics of Bayesian Methods	132
4.3 Involving Object Features in Classification	134
4.4 Bayesian Classification — a General Case	138
Multiple Features. Feature Vector	138

Bayes' Classification Rule for Multiclass Multifeature Objects	139
4.5 Statistical Classification Minimizing Risk	140
Bayesian Classification Minimizing the Probability of Error	143
Generalization of the Maximum Likelihood Classification	144
4.6 Decision Regions. Probabilities of Errors	145
4.7 Discriminant Functions	147
Gaussian Discriminant Function for Two Class Recognition	150
Quadratic and Linear Discriminant Derived from the Bayes Rule	150
Limitations of Bayesian Normal Discriminant	160
4.8 Estimation of Probability Densities	161
Parametric Methods of Probability Density Estimation	162
Non-Parametric Methods of Probability Density Estimation	165
Semi-Parametric Methods of Probability Density Estimation	174
Distance Between Probability Densities. Kullback-Leibler Distance	178
4.9 Probabilistic Neural Network (PNN)	179
Design of the Probabilistic Neural Network (PNN)	182
Probabilistic Neural Network with the Radial Gaussian Kernel	183
4.10 Constraints in Design	185
4.11 Summary	186
4.12 Exercises	186
References	189
 CHAPTER 5 EVOLUTIONARY COMPUTING	 193
5.1 Genetic Algorithms. Concept and Algorithmic Aspects	193
5.2 Fundamental Components of GAs	196
Encoding and Decoding	196
Selection Mechanism	197
Crossover and Mutation	198
Rule Encoding—a Simple Example	199
5.3 GA. Formal Definition of Genetic Algorithms	202
5.4 Schemata Theorem: a Conceptual Backbone of GAs	203
5.5 Genetic Computing. Further Enhancement	208
Floating Point Encoding	209
Selection Mechanisms	210
5.6 Exploration and Exploitation of the Search Space	212
5.7 Experimental Studies	213
Genetic Data Mining	215
Genetic Operators	216
Fitness Function	218
5.8 Classes of Evolutionary Computation	219
Evolution Strategies	219
Evolutionary Programming	220
Genetic Programming	221
5.9 Genetic Optimization of Rule-Based Description of Data: Pittsburgh and Michigan Approaches	224
5.10 Summary	225
5.11 Exercises	225

References	226
CHAPTER 6 MACHINE LEARNING	229
6.1 Introduction	229
Taxonomy of Machine Learning Algorithms	232
Why Machine Learning?	235
6.2 Introduction to Generation of Hypotheses	236
6.3 Overfitting	239
6.4 Rule Algorithms	241
6.5 Decision Tree Algorithms	249
Calculating Entropy and Information Gain	250
Growing Decision Trees and Rule Extraction	252
Extensions of ID Algorithms	254
6.6 Hybrid Algorithms	257
6.7 Discretization of Continuous-Valued Attributes	263
Equal Width Discretization	264
Equal Frequency Discretization	264
K-means Clustering Discretization	264
One-level Decision Tree Discretization	265
6.7.1 Information-Theoretic Discretization Methods	266
Partition	266
Quanta matrix	266
χ^2 Test	270
Maximum Entropy Discretization	272
Class-Attribute Interdependence Redundancy Discretization (CAIR)	273
Class-Attribute Interdependence Uncertainty and Redundancy Discretization(CAIUR)	274
6.8 Hypothesis Evaluation	275
Accuracy Test	276
Verification Test	277
6.9 Comparison of the Three Families of Algorithms	279
Discrete MONK's Data	279
Continuous IRIS Data	281
6.10 Machine Learning in Knowledge Discovery	283
6.11 Machine Learning and Rough Sets	285
6.12 Summary	286
6.13 Exercises	286
References	286
Appendix A6: Diagnosing Coronary Artery Disease (CAD)	289
Scintigraphic Images of the Heart	289
Data Preprocessing	290
Data Mining Methods Used	294
Results of Other Data Mining Methods on the CAD Data	300
References	307

CHAPTER 7 NEURAL NETWORKS	309
7.1 Introduction	309
Neuron Models	309
Neural Network Topologies	311
Learning Rules	311
Learning Algorithms	317
Why Neural Networks?	318
7.2 Radial Basis Function (RBF) Network	319
Basis Functions	322
Parameter Selection	325
Training for the Output Weights	329
Confidence Measures	332
7.3 RBF Networks in Knowledge Discovery	336
Rule-Based Indirect Interpretation of RBF Network	336
Fuzzy Context-based RBF Network	339
Using Neural Networks for Feature Ranking	342
7.4 Kohonen's Self Organizing Map(SOM)Network	344
Sammon's Projection	345
Self-Organizing Feature Maps	346
SOM Topology and Its Learning Rule	346
The Neighborhood Kernel	348
SOM Algorithm	351
Calibration of the SOM	352
Pragmatic Issues	353
7.5 Image Recognition Neural Network (IRNN)	357
Sensory Layer	357
Feature Aggregating Layer	360
Associative Layer	360
Learning in the Sensory and Feature Aggregating Layers	360
Training of the IRNN Network	365
7.6 Summary	367
7.7 Exercises	367
References	369
Appendix A7: Image Similarity(IS) Measure	372
Image Intensity Center	372
 CHAPTER 8 CLUSTERING	 375
8.1 Unsupervised Learning: a General Taxonomy and Related Algorithmic Aspects	375
Distance Functions and Proximity Matrices	377
8.2 Hierarchical Clustering	379
8.3 Objective Function—Based Clustering	382
8.4 Clustering Methods and Data Mining	384
Context-Oriented Fuzzy Clustering	385
The Context-Oriented Clustering Algorithm	387
Quantification of the Associations between Information Granules	396
Algorithmic Aspects of Context-Based Clustering	398

8.5 Hierarchical Clustering in Building Associations in the Data	403
8.6 Clustering under Partial Supervision in Data Mining	406
8.7 A Neural Realization of Similarity Between Patterns	412
8.8 Numerical Experiments	413
8.9 Summary	426
8.10 Exercises	426
References	428
CHAPTER 9 PREPROCESSING	431
9.1 Patterns and Features	431
Pattern Forming	432
Pattern Vectors (Patterns in a Vector Space)	433
Attributes (Features)	433
9.2 Preprocessing Operations	434
Data Types	434
Preprocessing and Its Goals	435
Sequence of Processing Steps	435
Basic Preprocessing of Raw Data	436
Feature Extraction from Raw Data	437
Pattern Forming for Defined Objects and Pattern Encoding	438
Forming an Information System	439
Basic Processing of an Information System	439
Feature Extraction/Transformations from an Information System	440
Feature Selection and Evaluation	441
9.3 Principal Component Analysis -- Feature Extraction and Reduction	441
Introduction	441
Modeling of Data as a Feature Extraction	443
Principal Component Analysis (PCA)	444
Dimensionality Reduction	450
9.4 Supervised Feature Reduction Based on Fisher's Linear Discriminant Analysis	451
Introduction	451
Two-Class Data and Fisher's Projection onto a Line	452
Multi-Class Fisher's Linear Discriminant Analysis and Transformation	455
9.5 Sequence of Karhunen-Loeve and Fisher's Linear Discriminant Projections	457
9.6 Feature Selection	459
Introduction	459
Optimal Feature Selection	460
General Paradigms of Optimal Feature Selection	460
Relevance of Features	461
Feature Selection Methods and Algorithms	462
Feature Selection Criteria	464
Open-loop Feature Selection Criteria	465
Closed-loop Feature Selection Criteria	468
Search Methods	469
Branch and Bound Method for Optimal Feature Selection	470

Feature Selection with Individual Feature Ranking	473
Suboptimal Sequential Forward Feature Selection	474
Feature Selection Based on Rough Sets	476
Overview of Feature Selection Methods	476
9.7 Numerical Experiments - Texture Image Classification	478
Singular Value Decomposition as a Feature Extraction from Images	479
Raw Data Preprocessing, Feature Extraction and Pattern Forming	481
Discussion of the Results	482
9.8 Summary	483
9.9 Exercises	483
References	486
INDEX	491

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.

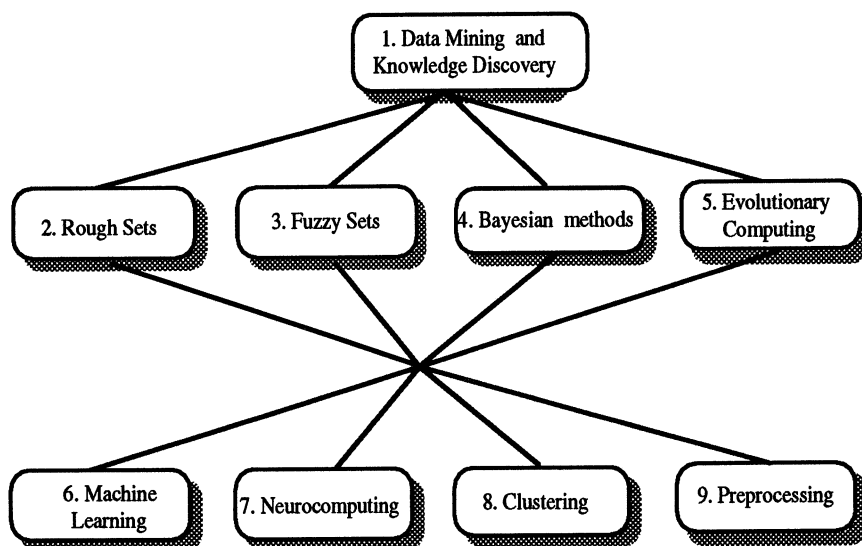


Figure 1. Data mining methods for knowledge discovery: a general roadmap

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