Data Mining and Knowledge Discovery Handbook

Second Edition

Oded Maimon · Lior Rokach
Editors

Data Mining and Knowledge Discovery Handbook

Second Edition

Springer

Editors

Prof. Oded Maimon
Tel Aviv University
Dept. Industrial Engineering
69978 Ramat Aviv
Israel
maimon@eng.tau.ac.il

Dr. Lior Rokach
Ben-Gurion University of the Negev
Dept. Information Systems
Engineering
84105 Beer-Sheva
Israel
liorrk@bgu.ac.il

ISBN 978-0-387-09822-7 e-ISBN 978-0-387-09823-4
DOI 10.1007/978-0-387-09823-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010931143

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my family
– Oded Maimon

To my parents Ines and Avraham
– Lior Rokach

Preface

Knowledge Discovery demonstrates intelligent computing at its best, and is the most desirable and interesting end-product of Information Technology. To be able to discover and to extract knowledge from data is a task that many researchers and practitioners are endeavoring to accomplish. There is a lot of hidden knowledge waiting to be discovered – this is the challenge created by today's abundance of data.

Knowledge Discovery in Databases (KDD) is the process of identifying valid, novel, useful, and understandable patterns from large datasets. Data Mining (DM) is the mathematical core of the KDD process, involving the inferring algorithms that explore the data, develop mathematical models and discover significant patterns (implicit or explicit) -which are the essence of useful knowledge. This detailed guide book covers in a succinct and orderly manner the methods one needs to master in order to pursue this complex and fascinating area.

Given the fast growing interest in the field, it is not surprising that a variety of methods are now available to researchers and practitioners. This handbook aims to organize all major concepts, theories, methodologies, trends, challenges and applications of Data Mining into a coherent and unified repository. This handbook provides researchers, scholars, students and professionals with a comprehensive, yet concise source of reference to Data Mining (and additional selected references for further studies).

The handbook consists of eight parts, each part consists of several chapters. The first seven parts present a complete description of different methods used throughout the KDD process. Each part describes the classic methods, as well as the extensions and novel methods developed recently. Along with the algorithmic description of each method, the reader is provided with an explanation of the circumstances in which this method is applicable, and the consequences and trade-offs incurred by using that method. The last part surveys software and tools available today.

The first part describes preprocessing methods, such as cleansing, dimension reduction, and discretization. The second part covers supervised methods, such as regression, decision trees, Bayesian networks, rule induction and support vector machines. The third part discusses unsupervised methods, such as clustering, association rules, link analysis and visualization. The fourth part covers soft computing methods and their application to Data Mining. This part includes chapters about fuzzy logic, neural networks, and evolutionary algorithms.

Parts five and six present supporting and advanced methods in Data Mining, such as statistical methods for Data Mining, logics for Data Mining, DM query languages, text mining, web mining, causal discovery, ensemble methods, and a great deal more. Part seven provides an in-depth description of Data Mining applications in various interdisciplinary industries, such as finance, marketing, medicine, biology, engineering, telecommunications, software, and security.

The motivation: Over the past few years we have presented and written several scientific papers and research books in this fascinating field. We have also developed successful methods for very large complex applications in industry, which are in operation in several enterprises. Thus, we have first hand experience in the needs of the KDD/DM community in research and practice. This handbook evolved from these experiences.

The first edition of the handbook, which was published five years ago, was extremely well received by the data mining research and development communities. The field of data mining has evolved in several aspects since the first edition. Advances occurred in areas, such as Multimedia Data Mining, Data Stream Mining, Spatio-temporal Data Mining, Sequences Analysis, Swarm Intelligence, Multi-label classification and privacy in data mining. In addition new applications and software tools become available. We received many requests to include the new advances in the field in a second edition of the handbook. About half of the book is new in this edition. This second edition aims to refresh the previous material in the fundamental areas, and to present new findings in the field. The new advances occurred mainly in three dimensions: new methods, new applications and new data types, which can be handled by new and modified advanced data mining methods.

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

Tel-Aviv, Israel *Oded Maimon*
Beer-Sheva, Israel *Lior Rokach*

April 2010

Contents

Part I Preprocessing Methods

Part II Supervised Methods

Part III Unsupervised Methods

Part IV Soft Computing Methods

Part V Supporting Methods

Part VI Advanced Methods

Part VII Applications

Part VIII Software

List of Contributors

Maria M. Abad
Software Engineering Department,
University of Granada, Spain

Ajith Abraham
Center of Excellence for Quantifiable
Quality of Service
Norwegian University of Science and
Technology,
Trondheim, Norway

Bruno Agard
Département de Mathmatiques et de
Génie Industriel,
École Polytechnique de Montréal,
Canada

Daniel Barbara
Department of Information and Software Engineering,
George Mason University, USA

Christopher D. Barko
Customer Analytics, Inc.

Irad Ben-Gal
Department of Industrial Engineering,
Tel-Aviv University, Israel

Moty Ben-Dov
School of Computing Science,
MDX University, London, UK.

Yoav Benjamini
Department of Statistics, Sackler
Faculty for Exact Sciences
Tel Aviv University, Israel

Richard A. Berk
Department of Statistics
UCLA, USA

Jean-Francois Boulicaut
INSA Lyon, France

Pavel Brazdil
Faculty of Economics,
University of Porto, Portugal

Abel Browarnik
Department of Industrial Engineering,
Tel-Aviv University, Israel

Christopher J.C. Burges
Microsoft Research, USA

Cory J. Butz
Department of Computer Science,
University of Regina, Canada

Nitesh V. Chawla
Department of Computer Science and
Engineering,
University of Notre Dame, USA

Ping Chen
Department of Computer and Mathematics Science,
University of Houston-Downtown, USA

Barak Chizi
Department of Industrial Engineering,
Tel-Aviv University, Israel

Shahar Cohen
Department of Industrial Engineering,
Tel-Aviv University, Israel

Antonio Congiusta
Dipartimento di Elettronica, Informatica e Sistemistica,
University of Calabria, Italy

Gautam Das
Computer Science and Engineering Department,
University of Texas, Arlington, USA

Swagatam Das
Department of Electronics and Telecommunication Engineering,
Jadavpur University, India.

Steve Donoho
Mantas, Inc. USA

Sašo Džeroski
Jožef Stefan Institute, Slovenia

Ronen Feldman
Department of Mathematics and Computer Science,
Bar-Ilan university, Israel

Eibe Frank
Department of Computer Science,
University of Waikato, New Zealand

Alex A. Freitas
Computing Laboratory,
University of Kent, UK

Johannes Fürnkranz
TU Darmstadt, Knowledge Engineering Group, Germany

Mohamed Medhat Gaber
Centre for Distributed Systems and Software Engineering
Monash University

Pierre Geurts
Department of Electrical Engineering and Computer Science,
University of Liège, Belgium

Christophe Giraud-Carrier
Department of Computer Science,
Brigham Young University, Utah, USA

Paolo Giudici
Faculty of Economics,
University of Pavia, Italy

Bart Goethals
Departement of Mathemati1cs and Computer Science,
University of Antwerp, Belgium

Jerzy W. Grzymala-Busse
Department of Electrical Engineering and Computer Science,
University of Kansas, USA

Witold J. Grzymala-Busse
FilterLogix Inc., USA

Dimitrios Gunopulos
Department of Computer Science and Engineering,
University of California at Riverside, USA

Petr Hájek
Institute of Computer Science,
Academy of Sciences of the Czech Republic

Maria Halkidi
Department of Computer Science and Engineering,
University of California at Riverside, USA

Mark Hall
Department of Computer Science,
University of Waikato, New Zealand

Howard J. Hamilton
Department of Computer Science,
University of Regina, Canada

Jiawei Han
Department of Computer Science,
University of Illinois, Urbana Champaign, USA

Geoffrey Holmes
Department of Computer Science,
University of Waikato, New Zealand

Frank Höppner
Department of Information Systems,
University of Applied Sciences Braunschweig/Wolfenbüttel, Germany

Yan Huang
Department of Computer Science,
University of Minnesota, USA

Sushil Jajodia
Center for Secure Information Systems,
George Mason University, USA

Ioannis Katakis
Dept. of Informatics, Aristotle University of Thessaloniki, 54124 Greece

Eamonn Keogh
Computer Science and Engineering Department,
University of California at Riverside, USA

Richard Kirkby
Department of Computer Science,
University of Waikato, New Zealand

Slava Kisilevich
University of Konstanz, Germany

Boris Kovalerchuk
Department of Computer Science,
Central Washington University, USA

Shonali Krishnaswamy
Centre for Distributed Systems and Software Engineering
Monash University

Andrew Kusiak
Department of Mechanical and Industrial Engineering,
The University of Iowa, USA

Nada Lavrač
Jožef Stefan Institute, Ljubljana, Slovenia
Nova Gorica Polytechnic, Nova Gorica, Slovenia

Moshe Leshno
Faculty of Management and Sackler Faculty of Medicine,
Tel Aviv University, Israel

Nissan Levin
Q-Ware Software Company, Israel

Tao Li
School of Computer Science,
Florida International University, USA

Churn-Jung Liau
Institute of Information Science,
Academia Sinica, Taiwan

Jessica Lin
Department of Computer Science and Engineering,
University of California at Riverside, USA

Tsau Y. Lin
Department of Computer Science,
San Jose State University, USA

Sheng Ma
Machine Learning for Systems
IBM T.J. Watson Research Center, USA

Oded Maimon
Department of Industrial Engineering,
Tel-Aviv University, Israel

Jonathan I. Maletic
Department of Computer Science,
Kent State University, USA

Florian Mansmann
University of Konstanz, Germany

Andrian Marcus
Department of Computer Science,
Wayne State University, USA

Cyrille Masson
INSA Lyon, France

Steve Moyle
Computing Laboratory,
Oxford University, UK

Mirco Nanni
University of Pisa,
Italy

Hamid R. Nemati
Information Systems and Operations
Management Department
Bryan School of Business and Economics
The University of North Carolina at
Greensboro, USA

Mitsunori Ogihara
Computer Science Department,
University of Rochester, USA

Nora Oikonomakou
Department of Informatics,
Athens University of Economics and Business (AUEB), Greece

Bernhard Pfahringer
Department of Computer Science,
University of Waikato, New Zealand

Marco F. Ramoni
Departments of Pediatrics and Medicine
Harvard University, USA

Chotirat Ann Ratanamahatana
Department of Computer Science and Engineering,
University of California at Riverside, USA

Yoram Reich
Center for Design Research,
Stanford University, Stanford, CA, USA

Salvatore Rinzivillo
Institute of Information Science
and Technologies,
Italy

Lior Rokach
Department of Information Systems Engineering
Ben-Gurion University of the Negev,
Israel

Noa Ruschin Rimini
Department of Industrial Engineering,
Tel-Aviv University, Israel

Sigal Sahar
Department of Computer Science,
Tel-Aviv University, Israel

Paola Sebastiani
Department of Biostatistics,
Boston University, USA

Richard S. Segall
Arkansas State University,
Department of Computer and Info.
Tech., Jonesboro, AR
72467-0130,USA

Shashi Shekhar
Institute of Technology,
University of Minnesota, USA

Armin Shmilovici
Department of Information Systems
Engineering,
Ben-Gurion University of the Negev,
Israel

Gautam B. Singh
Department of Computer Science and
Engineering,
Center for Bioinformatics, Oakland
University, USA

Anoop Singhal
Center for Secure Information Systems,
George Mason University, USA

Domenico Talia
Dipartimento di Elettronica, Informatica
e Sistemistica,
University of Calabria, Italy

Kurt Thearling
Vertex Business Services
Richardson, Texas, USA

Vicenç Torra
Institut d’Investigació en Intel·ligència
Artificial, Spain

Paolo Trunfio
Dipartimento di Elettronica, Informatica
e Sistemistica,
University of Calabria, Italy

Grigorios Tsoumakas
Dept. of Informatics, Aristotle University of Thessaloniki, 54124 Greece

Jiong Yang
Department of Electronic Engineering
and Computer Science,
Case Western Reserve University, USA

Ying Yang
School of Computer Science and
Software Engineering,
Monash University, Melbourne, Australia

Hong Yao
Department of Computer Science,
University of Regina, Canada

Philip S. Yu
IBM T. J. Watson Research Center,
USA

Michalis Vazirgiannis
Department of Informatics,
Athens University of Economics and
Business, Greece

Ricardo Vilalta
Department of Computer Science,
University of Houston, USA

Evgenii Vityaev
Institute of Mathematics,
Russian Academy of Sciences, Russia

Michail Vlachos
IBM T. J. Watson Research Center,
USA

Ioannis Vlahavas
Dept. of Informatics, Aristotle University of Thessaloniki, 54124 Greece

Haixun Wang
IBM T. J. Watson Research Center,
USA

Wei Wang
Department of Computer Science,
University of North Carolina at Chapel
Hill, USA

Geoffrey I. Webb
Faculty of Information Technology,
Monash University, Australia

Gary M. Weiss
Department of Computer and Information Science,
Fordham University, USA

Ian H. Witten
Department of Computer Science,
University of Waikato, New Zealand

Jacob Zahavi
The Wharton School,
University of Pennsylvania, USA

Arkady Zaslavsky
Centre for Distributed Systems and
Software Engineering
Monash University

Peter G. Zhang
Department of Managerial Sciences,
Georgia State University, USA

Pusheng Zhang
Department of Computer Science and
Engineering,
University of Minnesota, USA

Qingyu Zhang
Arkansas State University, Department
of
Computer and Info. Tech.,
Jonesboro, AR 72467-0130,USA

Ruofei Zhang
Yahoo!, Inc. Sunnyvale, CA 94089

Zhongfei (Mark) Zhang
SUNY Binghamton, NY 13902-6000

Blaž Zupan
Faculty of Computer and Information
Science,
University of Ljubljana, Slovenia

1

Introduction to 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

Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. *Data Mining* (DM) is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction.

The accessibility and abundance of data today makes Knowledge Discovery and Data Mining a matter of considerable importance and necessity. Given the recent growth of the field, it is not surprising that a wide variety of methods is now available to the researchers and practitioners. No one method is superior to others for all cases. The handbook of Data Mining and Knowledge Discovery from Data aims to organize all significant methods developed in the field into a coherent and unified catalog; presents performance evaluation approaches and techniques; and explains with cases and software tools the use of the different methods. The goals of this introductory chapter are to explain the KDD process, and to position DM within the information technology tiers. Research and development challenges for the next generation of the science of KDD and DM are also defined. The rationale, reasoning and organization of the handbook are presented in this chapter for helping the reader to navigate the extremely rich and detailed content provided in this handbook. In this chapter there are six sections followed by a brief discussion of the changes in the second edition.
1. The KDD Process 2. Taxonomy of Data Mining Methods 3. Data Mining within the Complete Decision Support System 4. KDD & DM Research Opportunities and Challenges 5. KDD & DM Trends 6. The Organization of the Handbook 7. New to This Edition
The special recent aspects of data availability that are promoting the rapid development of KDD and DM are the electronically readiness of data (though of different types and reliability). The internet and intranet fast development in particular pro-

O. Maimon, L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, 2nd ed.,
DOI 10.1007/978-0-387-09823-4_1, © Springer Science+Business Media, LLC 2010

mote data accessibility (as formatted or unformatted, voice or video, etc.). Methods that were developed before the Internet revolution considered smaller amounts of data with less variability in data types and reliability. Since the information age, the accumulation of data has become easier and less costly. It has been estimated that the amount of stored information doubles every 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, security and many other fields. The data availability is increasing exponentially, while the human processing level is almost constant. Thus the potential gap increases exponentially. This gap is the opportunity for the KDD\DM field, which therefore becomes increasingly important and necessary.

1.1 The KDD Process

The knowledge discovery process (Figure 1.1) is iterative and interactive, consisting of nine steps. Note that the process is iterative at each step, meaning that moving back to adjust 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 deeply understand the process and the different needs and possibilities in each step. Taxonomy for the Data Mining methods is helping in this process. It is presented in the next section.

The process starts with determining the KDD goals, and "ends" with the implementation of the discovered knowledge. As a result, changes would have to 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 transformation, 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 and tuning of this step. Having understood the KDD goals, the preprocessing of the data starts, as defined in the next three steps (note that some of the methods here are similar to Data Mining algorithms, but are used in the preprocessing context):
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

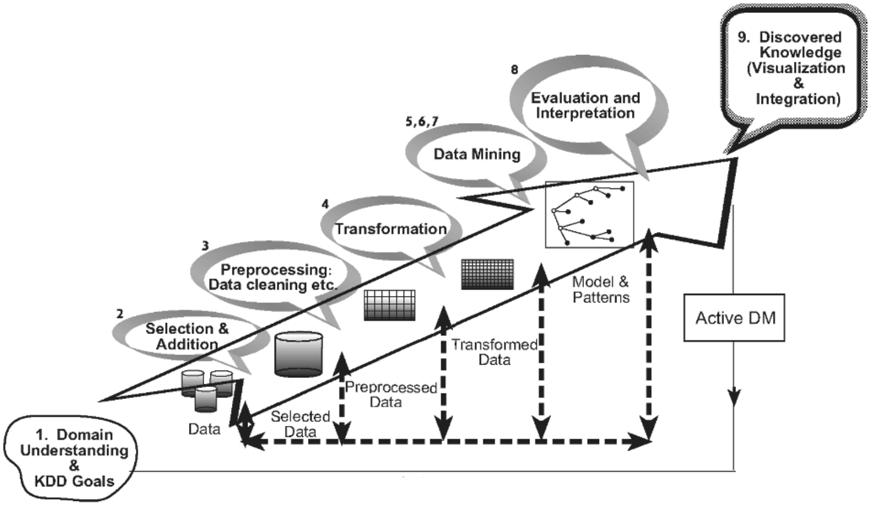
9. Discovered Knowledge (Visualization & Integration)
8
Evaluation and Interpretation
5,6,7
Data Mining
4
Transformation
3
Preprocessing: Data cleaning etc.
2
Selection & Addition
Data
Selected Data
Preprocessed Data
Transformed Data
Model & Patterns
Active DM
1. Domain Understanding & KDD Goals

Fig. 1.1. The Process of Knowledge Discovery in Databases.

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 that will be considered for the process. This process is very important because the Data Mining learns and discovers from the available data. This is the evidence base for constructing the models. If some important attributes are missing, then the entire study may fail. From success of the process it is good to consider as many as possible attribute at this stage. On the other hand, to collect, organize and operate complex data repositories is expensive, and there is a tradeoff with the opportunity for best understanding the phenomena. This tradeoff represents an aspect where the interactive and iterative aspect of the KDD is taking place. It starts with the best available data set and later expands and observes the effect in terms of knowledge discovery and modeling.

3. Preprocessing and cleansing. In this stage, data reliability is enhanced. It includes data clearing, such as handling missing values and removal of noise or outliers. Several methods are explained in the handbook, from doing nothing to becoming the major part (in terms of time consumed) of a KDD process in certain projects. It may involve complex statistical methods, or using specific Data Mining algorithm in this context. For example, if one suspects that a certain attribute is not reliable enough or has too many missing data, then this attribute could become the goal of a data mining supervised algorithm. A prediction model for this attribute will be developed, and then missing data can be predicted. The extension to which one pays attention to this level depends on many factors. In any case, studying these aspects is important and often revealing insight by itself, regarding enterprise information systems.

4. Data transformation. In this stage, the generation of better data for the data mining is prepared and developed. Methods here include dimension reduction (such as feature selection and extraction, and record sampling), and attribute transformation (such as discretization of numerical attributes and functional transformation). This step is often crucial for the success of the entire KDD project, but it is usually very project-specific. For example, in medical examinations, the quotient of attributes may often be the most important factor, and not each one by itself. In marketing, we may need to consider effects beyond our control as well as efforts and temporal issues (such as studying the effect of advertising accumulation). However, even if we do not use the right transformation at the beginning, we may obtain a surprising effect that hints to us about the transformation needed (in the next iteration). Thus the KDD process reflects upon itself and leads to an understanding of the transformation needed (like a concise knowledge of an expert in a certain field regarding key leading indicators). Having completed the above four steps, the following four steps are related to the Data Mining part, where the focus is on the algorithmic aspects employed for each project:
5. Choosing the appropriate Data Mining task. We are now ready to decide on which type of Data Mining to use, for example, classification, regression, or clustering. This mostly depends on the KDD goals, and also on the previous steps. There are two major goals in Data Mining: prediction and description. Prediction is often referred to as supervised Data Mining, while descriptive Data Mining includes the unsupervised and visualization aspects of Data Mining. Most data mining techniques are based on inductive learning, where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future cases. The strategy also takes into account the level of meta-learning for the particular set of available data.
6. Choosing the Data Mining algorithm. Having the strategy, we now decide on the tactics. This stage includes selecting the specific method to be used for searching patterns (including multiple inducers). For example, in considering precision versus understandability, the former is better with neural networks, while the latter is better with decision trees. For each strategy of meta-learning there are several possibilities of how it can be accomplished. Meta-learning focuses on explaining what causes a Data Mining algorithm to be successful or not in a particular problem. Thus, this approach attempts to understand the conditions under which a Data Mining algorithm is most appropriate. Each algorithm has parameters and tactics of learning (such as ten-fold cross-validation or another division for training and testing).
7. Employing the Data Mining algorithm. Finally the implementation of the Data Mining algorithm is reached. In this step we might need to employ the algorithm several times until a satisfied result is obtained, for instance by tuning the algorithm's control parameters, such as the minimum number of instances in a single leaf of a decision tree.
8. Evaluation. In this stage we evaluate and interpret the mined patterns (rules, reliability etc.), with respect to the goals defined in the first step. Here we consider

the preprocessing steps with respect to their effect on the Data Mining algorithm results (for example, adding features in Step 4, and repeating from there). This step focuses on the comprehensibility and usefulness of the induced model. In this step the discovered knowledge is also documented for further usage. The last step is the usage and overall feedback on the patterns and discovery results obtained by the Data Mining:

9. Using the discovered knowledge. We are now ready to incorporate the knowledge into another system for further action. The knowledge becomes active in the sense that we may make changes to the system and measure the effects. Actually the success of this step determines the effectiveness of the entire KDD process. There are many challenges in this step, such as loosing the "laboratory conditions" under which we have operated. For instance, the knowledge was discovered from a certain static snapshot (usually sample) of the data, but now the data becomes dynamic. Data structures may change (certain attributes become unavailable), and the data domain may be modified (such as, an attribute may have a value that was not assumed before).

1.2 Taxonomy of Data Mining Methods

There are many methods of Data Mining used for different purposes and goals. Taxonomy is called for to help in understanding the variety of methods, their interrelation and grouping. It is useful to distinguish between two main types of Data Mining: verification-oriented (the system verifies the user's hypothesis) and discovery-oriented (the system finds new rules and patterns autonomously). Figure 1.2 presents this taxonomy.

Discovery methods are those that automatically identify patterns in the data. The discovery method branch consists of prediction methods versus description methods. Descriptive methods are oriented to data interpretation, which focuses on understanding (by visualization for example) the way the underlying data relates to its parts. Prediction-oriented methods aim to automatically build a behavioral model, which obtains new and unseen samples and is able to predict values of one or more variables related to the sample. It also develops patterns, which form the discovered knowledge in a way which is understandable and easy to operate upon. Some prediction-oriented methods can also help provide understanding of the data.

Most of the discovery-oriented Data Mining techniques (quantitative in particular) are based on inductive learning, where a model is constructed, explicitly or implicitly, by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future unseen examples.

Verification methods, on the other hand, deal with the evaluation of a hypothesis proposed by an external source (like an expert etc.). These methods include the most common methods of traditional statistics, like goodness of fit test, tests of hypotheses (e.g., t-test of means), and analysis of variance (ANOVA). These methods are less associated with Data Mining than their discovery-oriented counterparts, because

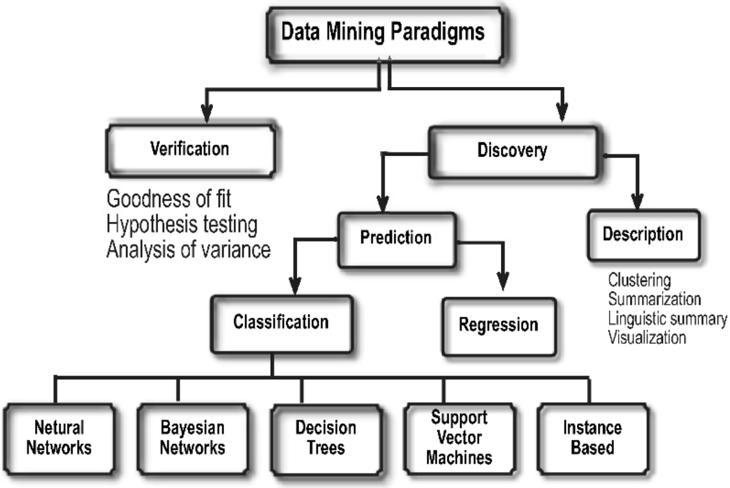

Fig. 1.2. Data Mining Taxonomy.

most Data Mining problems are concerned with discovering an hypothesis (out of a very large set of hypotheses), rather than testing a known one. Much of the focus of traditional statistical methods is on model estimation as opposed to one of the main objectives of Data Mining: model identification and construction, which is evidence based (though overlap occurs).

Another common terminology, used by the machine-learning community, refers to the prediction methods as supervised learning, as opposed to unsupervised learning. Unsupervised learning refers to modeling the distribution of instances in a typical, high-dimensional input space.

Unsupervised learning refers mostly to techniques that group instances without a prespecified, dependent attribute. Thus the term "unsupervised learning" covers only a portion of the description methods presented in Figure 1.2. For instance, it covers clustering methods but not visualization methods. Supervised methods are methods that attempt to discover the relationship between input attributes (sometimes called independent variables) and a target attribute sometimes referred to as a dependent variable). The relationship discovered is represented in a structure referred to as a model. Usually models describe and explain phenomena, which are hidden in the data set and can be used for predicting the value of the target attribute knowing the values of the input attributes. The supervised methods can be implemented on a variety of domains, such as marketing, finance and manufacturing. It is useful to distinguish between two main supervised models: classification models and regression models. The latter map the input space into a real-valued domain. For instance, a regressor can predict the demand for a certain product given its characteristics. On the other hand, classifiers map the input space into predefined classes. For example, classifiers can be used to classify mortgage consumers as good (fully payback the mortgage on time) and bad (delayed payback), or as many target classes as needed. There are many alternatives to represent classifiers. Typical examples include, support vector machines, decision trees, probabilistic summaries, or algebraic function.

1.3 Data Mining within the Complete Decision Support System

Data Mining methods are becoming part of integrated Information Technology (IT) software packages. Figure 1.3 illustrates the three tiers of the decision support aspect of IT. Starting from the data sources (such as operational databases, semi- and non-structured data and reports, Internet sites etc.), the first tier is the data warehouse, followed by OLAP (On Line Analytical Processing) servers and concluding with analysis tools, where Data Mining tools are the most advanced.

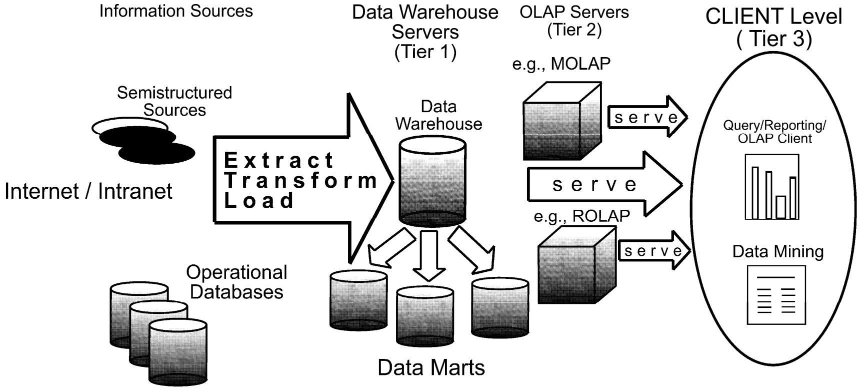

Fig. 1.3. The IT Decision Support Tiers.

The main advantage of the integrated approach is that the preprocessing steps are much easier and more convenient. Since this part is often the major burden for the KDD process (and can consumes most of the KDD project time), this industry trend is very important for expanding the use and utilization of Data Mining. However, the risk of the integrated IT approach comes from the fact that DM techniques are much more complex and intricate than OLAP, for example, so the users need to be trained appropriately.

This handbook shows the variety of strategies, techniques and evaluation measurements. We can naively distinguish among three levels of analysis. The simplest one is achieved by report generators (for example, presenting all claims that occurred because of a certain cause last year, such as car theft). We then proceed to OLAP multi-level analysis (for example presenting the ten towns where there was the highest increase of vehicle theft in the last month as compared to with the month before). Finally a complex analysis is carried out in discovering the patterns that predict car thefts in these cities, and what might occur if anti theft devices were installed. The latter is based on mathematical modeling of the phenomena, where the first two levels are ways of data aggregation and fast manipulation.

1.4 KDD and DM Research Opportunities and Challenges

Empirical comparison of the performance of different approaches and their variants in a wide range of application domains has shown that each performs best in some, but not all, domains. This phenomenon is known as the selective superiority problem, which means, in our case, that no induction algorithm can be the best in all possible domains. The reason is that each algorithm contains an explicit or implicit bias that leads it to prefer certain generalizations over others, and it will be successful only as long as this bias matches the characteristics of the application domain. Results have demonstrated the existence and correctness of this “no free lunch theorem”. If one inducer is better than another in some domains, then there are necessarily other domains in which this relationship is reversed. This implies in KDD that for a given problem a certain approach can yield more knowledge from the same data than other approaches.

In many application domains, the generalization error (on the overall domain, not just the one spanned in the given data set) of even the best methods is far above the training set, and the question of whether it can be improved, and if so how, is an open and important one. Part of the answer to this question is to determine the minimum error achievable by any classifier in the application domain (known as the optimal Bayes error). If existing classifiers do not reach this level, new approaches are needed. Although this problem has received considerable attention, no generally reliable method has so far been demonstrated. This is one of the challenges of the DM research – not only to solve it, but even to quantify and understand it better. Heuristic methods can then be compared absolutely and not just against each other.

A subset of this generalized study is the question of which inducer to use for a given problem. To be even more specific, the performance measure needs to be defined appropriately for each problem. Though there are some commonly accepted measures it is not enough. For example, if the analyst is looking for accuracy only, one solution is to try each one in turn, and by estimating the generalization error, to choose the one that appears to perform best. Another approach, known as multistrategy learning, attempts to combine two or more different paradigms in a single algorithm. The dilemma of which method to choose becomes even greater if other factors, such as comprehensibility are taken into consideration. For instance, for a specific domain, neural networks may outperform decision trees in accuracy. However from the comprehensibility aspect, decision trees are considered superior. In other words, in this case even if the researcher knows that neural network is more accurate, the dilemma of what methods to use still exists (or maybe to combine methods for their separate strength).

Induction is one of the central problems in many disciplines such as machine learning, pattern recognition, and statistics. However the feature that distinguishes Data Mining from traditional methods is its scalability to very large sets of varied types of input data. Scalability means working in an environment of high number of records, high dimensionality, and a high number of classes or heterogeneousness. Nevertheless, trying to discover knowledge in real life and large databases introduces time and memory problems. As large databases have become the norm in many fields (including astronomy, molecular biology, finance, marketing, health care, and many others), the use of Data Mining to discover patterns in them has become potentially very beneficial for the enterprise. Many companies are staking a large part of their future on these "Data Mining" applications, and turn to the research community for solutions to the fundamental problems they encounter. While a very large amount of available data used to be the dream of any data analyst, nowadays the synonym for "very large" has become "terabyte" or "pentabyte", a barely imaginable volume of information.

Information-intensive organizations (like telecom companies and financial institutions) are expected to accumulate several terabytes of raw data every one to two years. High dimensionality of the input (that is, the number of attributes) increases the size of the search space in an exponential manner (known as the "Curse of Dimensionality"), and thus increases the chance that the inducer will find spurious classifiers that in general are not valid. There are several approaches for dealing with a high number of records including: sampling methods, aggregation, massively parallel processing, and efficient storage methods.

1.5 KDD & DM Trends

This handbook covers the current state-of-the-art status of Data Mining. The field is still in its early stages in the sense that some basic methods are still being developed. The art expands but so does the understanding and the automation of the nine steps and their interrelation. For this to happen we need better characterization of the KDD problem spectrum and definition. The terms KDD and DM are not well-defined in terms of what methods they contain, what types of problem are best solved by these methods, and what results to expect. How are KDD\DM compared to statistics, machine learning, operations research, etc.? If subset or superset of the above fields? Or an extension\adaptation of them? Or a separate field by itself? In addition to the methods – which are the most promising fields of application and what is the vision KDD\DM brings to these fields? Certainly we already see the great results and achievements of KDD\DM, but we cannot estimate their results with respect to the potential of this field. All these basic analyses have to be studied and we see several trends for future research and implementation, including:

- Active DM – closing the loop, as in control theory, where changes to the system are made according to the KDD results and the full cycle starts again. Stability and controllability, which will be significantly different in these types of systems, need to be well-defined.

- Full taxonomy – for all the nine steps of the KDD process. We have shown a taxonomy for the DM methods, but a taxonomy is needed for each of the nine steps. Such a taxonomy will contain methods appropriate for each step (even the first one), and for the whole process as well.
- Meta-algorithms – algorithms that examine the characteristics of the data in order to determine the best methods, and parameters (including decompositions).
- Benefit analysis – to understand the effect of the potential KDD\DM results on the enterprise.
- Problem characteristics – analysis of the problem itself for its suitability to the KDD process.
- Mining complex objects of arbitrary type – Expanding Data Mining inference to include also data from pictures, voice, video, audio, etc. This will require adapting and developing new methods (for example, for comparing pictures using clustering and compression analysis).
- Temporal aspects - many data mining methods assume that discovered patterns are static. However, in practice patterns in the database evolve over time. This poses two important challenges. The first challenge is to detect when concept drift occurs. The second challenge is to keep the patterns up-to-date without inducing the patterns from scratch.
- Distributed Data Mining – The ability to seamlessly and effectively employ Data Mining methods on databases that are located in various sites. This problem is especially challenging when the data structures are heterogeneous rather than homogeneous.
- Expanding the knowledge base for the KDD process, including not only data but also extraction from known facts to principles (for example, extracting from a machine its principle, and thus being able to apply it in other situations).
- Expanding Data Mining reasoning to include creative solutions, not just the ones that appears in the data, but being able to combine solutions and generate another approach.

1.6 The Organization of the Handbook

This handbook is organized in eight parts. Starting with the KDD process, through to part six, the book presents a comprehensive but concise description of different methods used throughout the KDD process. Each part describes the classic methods as well as the extensions and novel methods developed recently. Along with the algorithmic description of each method, the reader is provided with an explanation of the circumstances in which this method is applicable and the consequences and the trade-offs of using the method including references for further readings. Part seven presents real-world case studies and how they can be solved. The last part surveys some software and tools available today. The first part is about preprocessing methods. This covers the preprocessing methods (Steps 3, 4 of the KDD process). The Data Mining methods are presented in the second part with the introduction and the very often-used supervised methods. The third part of the handbook considers the unsupervised methods. The fourth part is about methods termed soft computing, which include fuzzy logic, evolutionary algorithms, neural networks etc. Having established the foundation, we now proceed with supporting methods needed for Data Mining in the fifth part. The sixth part covers advanced methods like text mining and web mining. With all the methods described so far, the next section, the seventh, is concerned with applications for medicine, biology and manufacturing. The last and final part of this handbook deals with software tools. This part is not a complete survey of the software available, but rather a selected representative from different types of software packages that exist in today's market.

1.7 New to This Edition

Since the first edition that was published five years ago, the field of data mining has been evolved in the following aspects:

1.7.1 Mining Rich Data Formats

While in the past data mining methods could effectively analyze only flat tables, in recent years new mature techniques have been developed for mining rich data formats:

- Data Stream Mining - The conventional focus of data mining research was on mining resident data stored in large data repositories. The growth of technologies, such as wireless sensor networks, have contributed to the emergence of data streams. The distinctive characteristic of such data is that it is unbounded in terms of continuity of data generation. This form of data has been termed as data streams to express its owing nature. Mohamed Medhat Gaber, Arkady Zaslavsky, and Shonali Krishnaswamy present a review of the state of the art in mining data streams (Chapter 39). Clustering, classification, frequency counting, time series analysis techniques are been discussed. Different systems that use data stream mining techniques are also presented.
- Spatio-temporal - Spatio-temporal clustering is a process of grouping objects based on their spatial and temporal similarity. It is relatively new subfield of data mining, which gained high popularity especially in geographic information sciences due to the pervasiveness of all kinds of location-based or environmental devices that record position, time or/and environmental properties of an object or set of objects in real-time. As a consequence, different types and large amounts of spatio-temporal data became available and introduce new challenges to data analysis, which require novel approaches to knowledge discovery. Slava Kisilevich, Florian Mansmann, Mirco Nanni and Salvatore Rinzivillo provide a classification of different types of spatio-temporal data (Chapter 44). Then, they focus on one type of spatio-temporal clustering - trajectory clustering, provide an overview of the state-of-the-art approaches and methods of spatio-temporal clustering and finally present several scenarios in different application domains such as movement, cellular networks and environmental studies.

- Multimedia Data Mining - Zhongfei Mark Zhang and Ruofei Zhang present new methods for Multimedia Data Mining (Chapter 57). Multimedia data mining, as the name suggests, presumably is a combination of the two emerging areas: multimedia and data mining. Instead, the multimedia data mining research focuses on the theme of merging multimedia and data mining research together to exploit the synergy between the two areas to promote the understanding and to advance the development of the knowledge discovery multimedia data.

1.7.2 New Techniques

In this edition the following two new techniques are covered:

- In Chapter 23, Swagatam Das and Ajith Abraham present a family of bio-inspired algorithms, known as Swarm Intelligence (SI). SI has successfully been applied to a number of real world clustering problems. This chapter explores the role of SI in clustering different kinds of datasets. It also describes a new SI technique for partitioning a linearly non-separable dataset into an optimal number of clusters in the kernel- induced feature space. Computer simulations undertaken in this research have also been provided to demonstrate the effectiveness of the proposed algorithm.
- Multi-label classification - Most of the research in the field of supervised learning has been focused on single label tasks, where training instances are associated with a single label from a set of disjoint labels. However, Textual data, such as documents and web pages, are frequently annotated with more than a single label. In Chapter 34, Grigorios Tsoumakas, Loannis Katakis and Loannis Vlahavas review techniques for addressing multi-label classification task grouped into the two categories: i) *problem transformation*, and ii) *algorithm adaptation*. The first group of methods is algorithm independent. They transform the learning task into one or more single-label classification tasks, for which a large bibliography of learning algorithms exists. The second group of methods extends specific learning algorithms in order to handle multi-label data directly.
- Sequences Analysis - In Chapter 29, Noa Ruschin Rimini and Oded Maimon introduce a new visual analysis technique of sequences dataset using Iterated Function System (IFS). IFS produces a fractal representation of sequences. The proposed method offers an effective tool for visual detection of sequence patterns influencing a target attribute, and requires no understanding of mathematical or statistical algorithms. Moreover, it enables to detect sequence patterns of any length, without predefining the sequence pattern length.

1.7.3 New Application Domains

A new domain for KDD is the world of nanoparticles. Oded Maimon and Abel Browarnik present a smart repository system with text and data mining for this domain (Chapter 66). The impact of nanoparticles on health and the environment is a significant research subject, driving increasing interest from the scientific community, regulatory bodies and the general public. The growing body of knowledge in this area, consisting of scientific papers and other types of publications (such as surveys and whitepapers) emphasize the need for a methodology to alleviate the complexity of reviewing all the available information and discovering all the underlying facts, using data mining algorithms and methods. .

1.7.4 New Consideration

In Chapter 35, Vicenc Torra describes the main tools for privacy in data mining. He presents an overview of the tools for protecting data, and then focuses on protection procedures. Information loss and disclosure risk measures are also described.

1.7.5 Software

In Chapter 67, Zhang and Segall present selected commercial software for data mining, text mining, and web mining. The selected software are compared with their features and also applied to available data sets. Screen shots of each of the selected software are presented, as are conclusions and future directions.

1.7.6 Major Updates

Finally several chapters have been updated. Specifically, in Chapter 19, Alex Freitas presents a brief overview of EAs, focusing mainly on two kinds of EAs, viz. Genetic Algorithms (GAs) and Genetic Programming (GP). Then the chapter reviews the main concepts and principles used by EAs designed for solving several data mining tasks, namely: discovery of classification rules, clustering, attribute selection and attribute construction.

In Chapter 21, Peter Zhang provides an overview of neural network models and their applications to data mining tasks. He provides historical development of the field of neural networks and presents three important classes of neural models including feed forward multilayer networks, Hopfield networks, and Kohonen's self-organizing maps.

In Chapter 24, we discuss how fuzzy logic extends the envelope of the main data mining tasks: clustering, classification, regression and association rules. We begin by presenting a formulation of the data mining using fuzzy logic attributes. Then, for each task, we provide a survey of the main algorithms and a detailed description (i.e. pseudo-code) of the most popular algorithms.

References

Arbel, R. and Rokach, L., Classifier evaluation under limited resources, Pattern Recognition Letters, 27(14): 1619–1631, 2006, Elsevier.