PREFACE

The recent advent of effective and efficient computing and mass storage media, combined with a plethora of data recording devices, has resulted in the availability of unprecedented amounts of data. A few years ago we were talking about mega bytes to express the size of a database. Now people talk about giga bytes or even tera bytes. It is not a coincidence that the terms "*mega,*" "*giga,*" and "*tera*" (not to be confused with "*terra*" or earth in Latin) mean in Greek "*large,*" "*giant,*" and "*monster,*" respectively.

This situation has created many opportunities but also many challenges. The new field of data mining and knowledge discovery from databases is the most immediate result of this explosion of information and availability of cost effective computing power. Its ultimate goal is to offer methods for analyzing large amounts of data and extracting useful new knowledge embedded in such data. As K.C. Cole wrote in her seminal book *The Universe and the Teacup: The Mathematics of Truth and Beauty*, "*...* nature bestows her blessings buried in mountains of garbage."

Another anonymous author stated poetically that "today we are giants of information but dwarfs of new knowledge."

On the other hand, the principles that are behind most data mining methods are not new to modern science: the danger related with the excess of information and with its interpretation already alarmed the medieval philosopher William of Occam (Okham) and convinced him to state its famous "razor," *entia non sunt multiplicanda prater necessitatem (*plurality should not be assumed without necessity*)*. Data mining is thus not to be intended as a new approach to knowledge, but rather as a set of tools that make it possible to gain from observation of new complex phenomena the insight necessary to increase our knowledge.

Traditional statistical approaches cannot cope successfully with the heterogeneity of the data fields and also with the massive amounts of data available for analysis. Since there are many different goals in analyzing data and also different types of data, there are also different data mining and knowledge discovery methods, specifically designed to deal with data that are crisp, fuzzy, deterministic, stochastic, discrete, continuous, categorical, or any combination of the above. Sometimes the goal is just to use historic data to predict the behavior of a natural or artificial system; in other cases the goal is to extract easily understandable knowledge that can assist us to better understand the behavior of different types of systems, such as a mechanical apparatus, a complex electronic device, a weather system or the symptoms of an illness.

Thus, there is a real need to have methods which can extract new knowledge in a way that is easily verifiable and also easily understandable by a very wide array of domain experts. Such domain experts may not have the computational and mathematical expertise to fully understand how a data mining approach extracts new knowledge. However, they may easily comprehend newly extracted knowledge, if such knowledge can be expressed in an intuitive manner.

The present book contains a comprehensive compilation of methods that aim at deriving new knowledge in a way that is easily understood by a wide array of domain experts and end users. Thus, the focus is on discussing methods which are based on rules when they express new knowledge. The most typical form of such rules is a decision rule expressed as: IF <*some condition is true*> THEN <*another condition will also be true*>.

It presents the combined research experience of its 40 authors. This collective experience was gathered during a long search for methods capable of gleaning new knowledge from data. The last page of each chapter has a brief biographical statement about its contributors, who are world renowned experts from Australia, Belarus, Belgium, Brazil, China, Italy, Russia, Singapore, Spain, the United Kingdom, and the U.S.A.

This book provides a unique perspective into the core of data mining and knowledge discovery (DM&KD), combining many theoretical foundations for the behavior and capabilities of various DM&KD methods. It also presents a rich collection of examples, many of which come from real-life applications. A truly unique characteristic of this book is that almost all theoretical developments are accompanied by an extensive empirical analysis which often involved the solution of a very large number of simulated test problems. The results of these empirical analyses are tabulated, graphically depicted, and analyzed in depth. In this way, the theoretical and empirical analyses presented in this book are complementary to each other, so the reader can gain both a deep theoretical and practical insight of the covered subjects.

Another unique characteristic of this book is that at the end of each chapter there is a description of some possible research problems for future research. It also presents an extensive and updated bibliography and references of all the covered subjects. These are very valuable characteristics for people who wish to get involved with new research in DM&KD theory and applications.

Therefore, *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques* can provide a useful insight for people who are interested in obtaining a deep understanding of some of the most frequently used DM&KD methods: it can be used as a textbook for senior undergraduate or graduate courses in data mining in engineering, computer science, and business schools. It can also provide a panoramic and

systematic exposure of related methods and problems to researchers. Finally, it can become a valuable guidance for practitioners who wish to take a more effective and critical approach to the solution of real-life DM&KD problems.

The arrangement of the chapters follows a natural exposition of the main subjects in rule induction for DM&KD theory and practice. All chapters are intended to be self-contained, each providing the necessary background and definitions for its comprehension.

The first chapter is written by Dr. Zakrevskij, a member of the National Academy of Science of Belarus. The algorithms described in this chapter deal with the solution of some key data mining and pattern recognition problems. The proposed approaches are based on Boolean logic. More specifically, these approaches are based on inductive inference when solving data mining problems, while for solving pattern recognition problems they are based on deductive inference. This chapter also discusses some general schemes for representing real world environments. Such environments use binary (Boolean) variables to describe entities of interest (observations grouped into different classification classes) and also more abstract concepts, such as the classification classes. Next, the extracted new knowledge can easily be described in terms of these binary variables.

The second chapter is written by Professor Triantaphyllou, one of the editors of this book. It describes his research in inferring classification rules in the form of *if—then* type of decision rules. As data one considers binary observations which are grouped into disjoint classes. Such binary data may be derived from continuous observations when one performs certain data transformations. The algorithmic developments are based on some optimization models, formulated in terms of special forms of the well-known set covering problem and solved by branch-and-bound approaches. Given certain performance metrics, these algorithms can be optimal or semi-optimal and also easily scalable. Thus, their complexities can be NP-complete or just polynomial on the size of the input data. This is done by employing a graph theoretic approach. Both theoretical and empirical results are described. Some empirical results indicate the high potential of these methods.

Chapter 3 is written by Dr. Naidenova, a distinguished member of the Military Medical Academy in Saint Petersburg, Russia. This chapter discusses some methods for solving data mining and pattern recognition problems. These methods are based on common sense reasoning. For the data mining problems an approach is described for inferring implicative logical rules from observations grouped into different classes. Other related developments deal with the problems of incremental and non-incremental

learning and also with problems related to the design of good diagnostic tests.

Chapter 4 is written by Professors Torvik and Triantaphyllou. This chapter deals with the use of monotonicity for solving certain types of interesting data mining problems. Roughly speaking, monotonicity in the data means that the dependent variable (or class indicator) tends to point to a certain value when the value of an independent variable (or variables) increases (or decreases). It seems like many real-life phenomena exhibit, to a certain degree, this kind of property. From the algorithmic point of view, the utilization of the monotonicity property offers many exciting advantages in the development of optimal or semi-optimal data mining algorithms and also for solving many complex real life data mining and knowledge discovery problems. Often the latter is possible with the use of nested monotone Boolean functions and also with the use of hierarchically organized monotone Boolean functions. Some potential applications are discussed as well.

Chapter 5 is written by Drs. Felici and Sun and Professor Truemper. The problem considered in this chapter is how to analyze training data (i.e., observations about the behavior of a system of interest) and infer a Boolean function which accurately classifies observations grouped into two classes. A special sub-sampling technique is proposed to enhance the quality of the inferred rules and to evaluate and control their precision. This technique is based on a voting scheme that combines the formulas obtained from different samples of the training data to determine the class membership of new data points. Probabilistic considerations add to this scheme the capability of predicting, with a high level of precision, the error in recognizing a new data point by using only the information contained in the training data. Such logic functions can next easily be translated into classification rules for the construction of intelligent systems. This chapter presents many theoretical issues related to this rule inference problem and also discusses many application possibilities.

The sixth chapter is written by Dr. Felici, also an editor of this book, together with Professor de Angelis and Dr. Mancinelli. It treats the problems of feature selection arising in the solution of data mining problems. When the number of variables, or features, that describe the data to be analyzed is large, it may be necessary to apply these techniques to select only those features that are relevant to the purpose of the data mining application of interest, while discarding those that are redundant or nonsignificant. It also provides an overview of the literature. This overview outlines the main components of a feature selection procedure, and then it provides several examples of the two main approaches to this problem. These two approaches are: Filter methods and Wrapper methods. The

authors propose a Filter method based on a modification of a well-known graph theoretical model, the *k-lightest subgraph* problem. The model is then applied to a number of test instances to evaluate its performance, and also is applied to a real application. This application is concerned with the use of a logic data miner to a database derived from the questionnaires of a survey on urban mobility.

Chapter 7 is written by Bartnikowski, Granberry, Mugan, and Professor Truemper. A common problem with almost all methods that infer logical rules from data is that the data must be in binary form. The problem now is how to represent data that initially have rational and/or nominal values in terms of data that are defined on binary variables. Such a transformation, if not properly directed, may lead to an explosion of the dimensions of the data without providing additional information useful to the data mining task under consideration. For this reason, the authors propose a new procedure to map rational and nominal values into logic variables. This procedure tries to maximize the information which is useful for the final objective while containing the dimensions of the new space. Once the original data are transformed into binary data, next various data mining techniques may be employed.

Chapter 8 is written by Professor Kusiak, and deals with a critical issue in many data mining and knowledge discovery applications: how to best describe the information embedded in the observations. Determining the features of the data is the main focus of the emerging "data farming" discipline. This chapter presents the basic notions of this discipline and also uses a number of illustrative examples.

Chapter 9 is written by Dr. Orsenigo and Professor Vercellis and describes an approach for inferring rules in terms of decision trees for classification and prediction accuracy. This decision tree inference approach is based on a particular type of support vector machines, called *discrete* support vector machines. Their characteristic is that they aim at minimizing the number of misclassified instances rather than the total misclassification distance. Such a problem is modeled as a mixed integer programming problem. Good solutions are found by a scheme based on a sequence of linear programming problems (LP). Some computational results demonstrate the high potential of this approach.

Chapter 10 is written by Dr. Lee and Professor Olafsson, and presents the development of a classification approach which is based on the induction of top-down decision trees. It studies multi-attribute decision tree induction and methods for improving their accuracy and simplicity. The authors also discuss a recently proposed algorithm which uses conjunctive and disjunctive combinations of two attributes for induction of better

decision trees. Their method is called SODI, for second order decision tree induction. They also present some promising empirical results.

Chapter 11 is written by Drs. Zhai, Kho, and Fok. This chapter studies rule inference problems from observations grouped into different classes when the data are imprecise and/or incomplete. A number of related approaches, such as fuzzy logic and the Dempster-Shafer theory of belief functions are described. However, the main developments in this chapter are based on rough set theory, a powerful modeling approach. Some illustrative applications are also described.

Chapter 12 is written by Professors Noda and Freitas. This chapter studies the inference of prediction rules which are not only accurate and comprehensible, but also interesting. By "interesting" the authors mean that the rules are surprising. Thus they offer some quantitative measures to capture this notion of interestingness. This measure is used to guide a search which considers both the prediction accuracy and the degree of interestingness of candidate rules. A specially tailored genetic algorithm, according to some empirical results provided in this chapter, seems to be highly effective in discovering interesting rules or "knowledge nuggets."

Chapter 13 is written by Drs. Kirley, Abbass, and McKay. Here they discuss the development of classifier systems which are based on genetic algorithms (GAs). Different algorithmic issues are explored in order to determine what GA characteristics are important. More precisely, this chapter studies Pitt-style evolutionary classifier systems. Different performance measures and different computing platforms (sequential and parallel) are explored as well.

Chapter 14 is written by Professors G. Chen, Wei, and Kerre. This chapter studies different settings in inferring association rules from databases. Such rules can capture interesting patterns in the way items of interest occur simultaneously in database records. Thus, they have found a wide application in analyzing consumer market behaviors. This chapter pays particular attention to the presence of impression or fuzziness in the description of the pertinent data. Besides typical association rules the authors also study the inference of functional dependencies and pattern associations.

Chapter 15 is written by Professor Liao, and contains a rather wide literature review of methods that mine decision rules from fuzzy data. It considers a variety of fuzzy data mining approaches such as fuzzy clustering, fuzzy-neural networks, and fuzzy decision trees. It also considers genetic algorithms, and traditional neural networks. The results are very effectively summarized in extensive tables, where more than 100 literature references are classified and compared on the basis of some particularly interesting parameters.

Chapter 16 deals with the case of analyzing data for medical applications. Many data mining and knowledge discovery approaches for medical applications require the use of certain data preprocessing and feature extraction techniques. Such techniques may, for instance, enhance the contrast in medical images. Another related issue is how to process patient records that are based on textual and multimedia information. All the above and other related issues are discussed in detail in this chapter, which is written by Professors Elmaghraby, Kantardzic, and Wachowiak.

Chapter 17 is written by Professors Al-Mubaid and Truemper. It deals with a very interesting problem in developing a spelling checker for word processing systems. Traditional spelling checkers are very good at identifying spelling errors. However, they are weak at identifying the misuse of words that sound similar but are spelled differently. An example is the words "sight" versus "site." Unfortunately, most systems cannot identify such errors and oftentimes text may contain embarrassing errors. The approach proposed in this chapter offers high hope for successfully dealing with this computational challenge. The proposed method is based on the use of some training data (text documents) which are analyzed only once to extract information about the use of certain words. Such information is extracted in the form of Boolean functions. Once this phase is executed, new documents can be checked for this kind of spelling errors rather quickly.

Chapter 18 is written by Professors J. Chen, Kraft, Martin-Bautista, and Vila. These authors present their research findings on the induction and inference of fuzzy rules for textual information retrieval. Applications of these findings can be used when dealing with web related problems such as those where one performs a search based on some keywords of interest to the user. The empirical results provided at the end of this chapter provide solid experimental evidence of the high potential of these approaches.

Chapter 19 is written by Dr. Judson from the U.S. Census Bureau, and treats a critical problem in databases: how to link records of one database with records of another database. This is known as the record linkage problem (or RLP). When properly solved, such a linking offers the possibility to better describe entities of interest for which we have fragmented information stored in different databases. Solving the RLP involves solving a classification problem. Classes describe the "match" and "do not match" decision to be made when one considers a pair of records taken from each one of two databases. In turn, this classification problem is solved by means of Bayesian logistic regression and also by the application of the well-known Fellegi-Sunter model for the record linkage problem.

Finally, chapter 20 provides some reflections about the future of some data mining and knowledge discovery areas. These areas are in web mining, text mining, visual mining, and distributed mining. As is often the

case with the advent of highly promising computer technologies, at the beginning there is lots of hype in the expectations of what such methods can accomplish and what they cannot. This is also true with the data mining and knowledge discovery fields, and especially with methods which are based on the inference of rules. This chapter is written by Wang, Zhu, Felici, and Triantaphyllou.

If one wishes to summarize the potential of these methods in one sentence, then most experts would agree that, despite some promotional hype, data mining and knowledge discovery methods will become even more important in the future as data storage, acquisition, and processing hardware systems become more cost effective and efficient. In other words, one may assert that data mining will establish itself as a necessity for all those who, for science, for business, or for the most various reasons, desire to increase their knowledge by discovering new opportunities hidden deeply in vast amounts of seemingly confusing data.