# Preface

There is already a plethora of books on data mining. So, what is new with this book? The answer is in its unique perspective in studying a series of interconnected key data mining and knowledge discovery problems both in depth and also in connection with other related topics and doing so in a way that stimulates the quest for more advancements in the future. This book is related to another book titled *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques* (published by Springer in the summer of 2006), which was co-edited by the author. The chapters of the edited book were written by 40 authors and co-authors from 20 countries and, in general, they are related to rule induction methods.

Although there are many approaches to data mining and knowledge discovery (DM&KD), the focus of this monograph is on the development and use of some novel mathematical logic methods as they have been pioneered by the author of this book and his research associates in the last 20 years. The author started the research that led to this publication in the early 1980s, when he was a graduate student at the Pennsylvania State University.

During this experience he has witnessed the amazing explosion in the development of effective and efficient computing and mass storage media. At the same time, a vast number of ubiquitous devices are selecting data on almost any aspect of modern life. The above developments create an unprecedented challenge to extract useful information from such vast amounts of data. Just a few years ago people were talking about megabytes to express the size of a huge database. Today people talk about gigabytes or even terabytes. 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.

The above situation has created many opportunities but many new and tough challenges too. The emerging field of data mining and knowledge discovery is the most immediate result of this extraordinary explosion on information and availability of cost-effective computing power. The ultimate goal of this new field 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." An anonymous author expressed a closely related concept by stating poetically that "today we are giants of information but dwarfs of new knowledge."

On the other hand, the principles that are behind many 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 (also known as Okham) and motivated him to state his famous "razor": *entia non sunt multiplicanda praeter necessitatem* (entities must not be multiplied (i.e., become more complex) beyond necessity). Even older is the story in the Bible of the Tower of Babel in which people were overwhelmed by new and ultraspecialized knowledge and eventually lost control of the most ambitious project of that time.

People dealt with data mining problems when they first tried to use past experience in order to predict or interpret new phenomena. Such challenges always existed when people tried to predict the weather, crop production, market conditions, and the behavior of key political figures, just to name a few examples. In this sense, the field of data mining and knowledge discovery is as old as humankind.

Traditional statistical approaches cannot cope successfully with the heterogeneity of the data fields and also with the massive amounts of data available today 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 to just 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 an illness.

Thus, there is a 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 who 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 methods described in this book offer just this opportunity. This book presents methods that deal with key data mining and knowledge discovery issues in an intuitive manner and in a natural sequence. These methods are based on mathematical logic. Such methods derive new knowledge in a way that can be easily understood and interpreted by a wide array of domain experts and end users. Thus, the focus is on discussing methods which are based on Boolean functions; which can then easily be transformed into rules when they express new knowledge. The most typical form of such rules is a decision rule of the form: IF ⟨*some condition(s) is (are) true*⟩ THEN ⟨*another condition should also be true*⟩.

Thus, this book provides a unique perspective into the essence of some fundamental data mining and knowledge discovery issues. It discusses the theoretical foundations of the capabilities of the methods described in this book. It also presents a wide collection of illustrative 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 involves 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 comprehensive and 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 this field.

Therefore, the book *Data Mining and Knowledge Discovery via Logic-Based Methods: Theory, Algorithms, and Applications* can provide a valuable insight for people who are interested in obtaining a deep understanding of some of the most frequently encountered data mining and knowledge discovery challenges. This book 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 guide for practitioners who wish to take a more effective and critical approach to the solution of real-life data mining and knowledge discovery problems.

The philosophy followed on the development of the subjects covered in this book was first to present and define the subject of interest in that chapter and do so in a way that motivates the reader. Next, the following three key aspects were considered for each subject: (i) a discussion of the related theory, (ii) a presentation of the required algorithms, and (iii) a discussion of applications. This was done in a way such that progress in any one of these three aspects would motivate progress in the other two aspects. For instance, theoretical advances make it possible to discover and implement new algorithms. Next, these algorithms can be used to address certain applications that could not be addressed before. Similarly, the need to handle certain real-life applications provides the motivation to develop new theories which in turn may result in new algorithms and so on. That is, these three key aspects are parts of a continuous closed loop in which any one of these three aspects feeds the other two.

Thus, this book deals with the pertinent theories, algorithms, and applications as a closed loop. This is reflected on the organization of each chapter but also on the organization of the entire book, which is comprised of two sections. The sections are titled "Part I: Algorithmic Issues" and "Part II: Application Issues." The first section focuses more on the development of some new and fundamental algorithms along with the related theory while the second section focuses on some select applications and case studies along with the associated algorithms and theoretical aspects. This is also shown in the Contents.

The arrangement of the chapters follows a natural exposition of the main subjects in rule induction for DM&KD theory and practice. Part I ("Algorithmic Issues") starts with the first chapter, which discusses the intuitive appeal of the main data

mining and knowledge discovery problems discussed throughout this monograph. It pays extra attention to the reasons that lead to formulate some of these problems as optimization problems since one always needs to keep control on the size (i.e., for size minimization) of the extracted new rules or when one tries to gain a deeper understanding of the system of interest by issuing a small number of new queries (i.e., for query minimization).

The second and third chapters present some sophisticated branch-and-bound algorithms for extracting a pattern (in the form of a compact Boolean function) from collections of observations grouped into two disjoint classes. The fourth chapter presents some fast heuristics for the same problem.

The fifth chapter studies the problem of guided learning. That is, now the analyst has the option to decide the composition of the observation to send to an expert or "oracle" for the determination of its class membership. Apparently, the goal now is to gain a good understanding of the system of interest by issuing a small number of inquiries of the previous type.

A related problem is studied in the sixth chapter. Now it is assumed that the analyst has two sets of examples (observations) and a Boolean function that is inferred from these examples. Furthermore, it is assumed that the analyst has a new example that invalidates this Boolean function. Thus, the problem is how to modify the Boolean function such that it satisfies all the requirements of the available examples plus the new example. This is known as the incremental learning problem.

Chapter 7 presents an intriguing duality relationship which exists between Boolean functions expressed in CNF (conjunctive normal form) and DNF (disjunctive normal form), which are inferred from examples. This dual relationship could be used in solving large-scale inference problems, in addition to other algorithmic advantages.

The chapter that follows describes a graph theoretic approach for decomposing large-scale data mining problems. This approach is based on the construction of a special graph, called the *rejectability graph*, from two collections of data. Then certain characteristics of this graph, such as its minimum clique cover, can lead to some intuitive and very powerful decomposition strategies.

Part II ("Application Issues") begins with Chapter 9. This chapter presents an intriguing problem related to any model (and not only those based on logic methods) inferred from grouped observations. This is the problem of the reliability of the model and it is associated with both the number of the training data (sampled observations grouped into two disjoint classes) and also the nature of these data. It is argued that many model inference methods today may derive models that cannot guarantee the reliability of their predictions/classifications. This chapter prepares the basic arguments for studying a potentially very critical type of Boolean functions known as *monotone* Boolean functions.

The problems of inferring a monotone Boolean function from inquiries to an expert ("oracle"), along with some key mathematical properties and some application issues are discussed in Chapters 10 and 11. Although this type of Boolean functions has been known in the literature for some time, it was the author of this book along with some of his key research associates who made some intriguing contributions

to this part of the literature in recent years. Furthermore, Chapter 11 describes some key problems in assessing the effectiveness of data mining and knowledge discovery models (and not only for those which are based on logic). These issues are referred to as the "three major illusions" in evaluating the accuracy of such models. There it is shown that many models which are considered as highly successful, in reality may even be totally useless when one studies their accuracy in depth.

Chapter 12 presents how some of the previous methods for inferring a Boolean function from observations can be used (after some modifications) to extract what is known in the literature as association rules. Traditional methods suffer the problem of extracting an overwhelming number of association rules and they are doing so in exponential time. The new methods discussed in this chapter are based on some fast (of polynomial time) heuristics that can derive a compact set of association rules.

Chapter 13 presents some new methods for analyzing and categorizing text documents. Since the Web has made possible the availability of immense textual (and not only) information easily accessible to anyone with access to it, such methods are expected to attract even more interest in the immediate future.

Chapters 14, 15, and 16 discuss some real-life case studies. Chapter 14 discusses the analysis of some real-life EMG (electromyography) signals for predicting muscle fatigue. The same chapter also presents a comparative study which indicates that the proposed logic-based methods are superior to some of the traditional methods used for this kind of analysis.

Chapter 15 presents some real-life data gathered from the analysis of cases suspected of breast cancer. Next these data are transformed into equivalent binary data and then some diagnostic rules (in the form of compact Boolean functions) are extracted by using the methods discussed in earlier chapters. These rules are next presented in the form of IF-THEN logical expressions (diagnostic rules).

Chapter 16 presents a combination of some of the proposed logic methods with fuzzy logic. This is done in order to objectively capture fuzzy data that may play a key role in many data mining and knowledge discovery applications. The proposed new method is demonstrated in characterizing breast lesions in digital mammography as lobular or microlobular. Such information is highly significant in analyzing medical data for breast cancer diagnosis.

The last chapter presents some concluding remarks. Furthermore, it presents twelve different areas that are most likely to experience high interest for future research efforts in the field of data mining and knowledge discovery.

All the above chapters make clear that methods based on mathematical logic already play an important role in data mining and knowledge discovery. Furthermore, such methods are almost guaranteed to play an even more important role in the near future as such problems increase both in complexity and in size.

*Evangelos Triantaphyllou*
Baton Rouge, LA
April 2010