Preface

The advancement in data collection, storage, and distribution technologies has far outpaced computational advances in techniques for analyzing and understanding data. This encourages researchers and practitioners to develop a new generation of tools and techniques for data mining (DM) and for knowledge discovery in databases (KDD). KDD is a broad area that integrates concepts and methods from several disciplines including the fields of statistics, databases, artificial intelligence, machine learning, pattern recognition, machine discovery, uncertainty modeling, data visualization, high performance computing, optimization, management information systems, and knowledge-based systems.

KDD is a multistep iterative process. The preparatory steps of KDD include data selection and/or sampling, preprocessing and transformation of data for the subsequent steps of the process. Data mining is the next step in the KDD process. Data mining algorithms are used to discover patterns, clusters and models from data. The outcomes of the data mining algorithms are then rendered into operational forms that are easy for people to visualize and understand.

The data mining part of KDD usually uses a model and search based algorithm to find patterns and models of interests. The commonly used techniques are decision trees, genetic programming, neural networks, inductive logic programming, rough sets, Bayesian statistics, optimisation and other approaches. That means, heuristic and optimisation have a major role to play in data mining and knowledge discovery. However, most data mining work resulting from the application of heuristic and optimisation techniques has been reported in a scattered fashion in a wide variety of different journals and conference proceedings. As such, different journal and conference publications tend to focus on a very special and narrow topic. It is high time that an archival book series publishes a special volume which provides critical reviews of the state-of-art applications of heuristic and optimisation techniques associated with data mining and KDD problems. This volume aims at filling in the gap in the current literature.

This special volume consists of open-solicited and invited chapters written by leading researchers in the field. All papers were peer reviewed by at least two recognised reviewers. The book covers the foundation as well as the practical side of data mining and knowledge discovery.

This book contains 15 chapters, which can be categorized into the following five sections:

- Section 1: Introduction
- Section 2: Search and Optimization
- Section 3: Statistics and Data Mining
- Section 4: Neural Networks and Data Mining
- Section 5: Applications

In the first chapter, an introduction to data mining and KDD, and the steps of KDD are briefly presented. The DM tasks and tools are also provided in this chapter. The role of heuristic and optimisation techniques in KDD are also discussed.

Section 2 contains Chapters 2 to 6. Chapter 2 presents an algorithm for feature selection, which is based on a conventional optimization technique. The effectiveness of the proposed algorithm is tested by applying it to a number of publicly available real-world databases. Chapter 3 reports the results obtained from a series of studies on cost-sensitive classification using decision trees, boosting algorithms, and MetaCost which is a recently proposed procedure that converts an error-based algorithm into a cost-sensitive algorithm. The studies give rise to new variants of algorithms designed for cost-sensitive classification, and provide insight into the strengths and weaknesses of the algorithms. Chapter 4 presents an optimization problem that addresses the selection of a combination of several classifiers such as boosting, bagging and stacking. This is followed by the discussion of heuristic search techniques, in particular, genetic algorithms applied to automatically obtain the ideal combination of learning methods for the stacking system. Chapter 5 examines the use of the Component Object Model (COM) in the design of search engines for knowledge discovery and data mining using modern heuristic techniques and how adopting this approach benefits the design of a commercial toolkit. The chapter also describes how search engines have been implemented as COM objects and how the representation and problem components have been created to solve rule induction problems in data mining. Chapter 6 discusses the possibility of applying the logical combinatorial pattern recognition (LCPR) tools to the clustering of large and very large mixed incomplete data (MID) sets. This research is directed towards the application of methods, techniques and in general, the philosophy of the LCPR to the solution of supervised and unsupervised classification problems. In this chapter, the clustering algorithms GLC, DGLC, and GLC+ are introduced.

Chapters 7 to 9 comprise Section 3. Chapter 7 introduces the Bayes' Theorem and discusses the applicability of the Bayesian framework to three traditional statistical and/or machine learning examples: a simple probability experiment involving coin-tossing, Bayesian linear regression and Bayesian neural network learning. Some of the problems associated with the practical aspects of the implementation of Bayesian learning are then detailed, followed by the introduction of various software that is freely available on the Internet. The advantages of the Bayesian approach to learning and inference, its impact on diverse scientific fields and its present applications are subsequently identified. Chapter 8 addresses the question of how to decide how large a sample is necessary in order to apply a particular data mining procedure to a given data set. A brief review of the main results of basic sampling theory is followed by a detailed consideration and comparison of the impact of simple random sample size on two well-known data mining procedures: naïve Bayes classifiers and decision tree induction. The chapter also introduces a more sophisticated form of sampling, disproportionate stratification, and shows how it may be used to make much more effective use of limited processing resources. Chapter 9 shows how the Gamma test can be used in the construction of predictive models and classifiers for numerical data. In doing so, the chapter also demonstrates the application of this technique to feature selection and to the selection of the embedding dimension when dealing with a time series.

Section 4 consists of Chapters 10 and 11. Neural networks are commonly used for prediction and classification when data sets are large. They have a major advantage over conventional statistical tools in that it is not necessary to assume any mathematical form for the functional relationship between the variables. However, they also have a few associated problems, like the risk of over-parametrization in the absence of P-values, the lack of appropriate diagnostic tools and the difficulties associated with model interpretation. These problems are particularly pertinent in the case of small data sets.

Chapter 10 investigates these problems from a statistical perspective in the context of typical market research data. Chapter 11 proposes an efficient on-line learning method called adaptive natural gradient learning. It can solve the plateau problems and can successfully be applied to learning involving large data sets.

The last section presents four application chapters, Chapters 12 to 15. Chapter 12 introduces rough clustering, a technique based on a simple extension of rough set theory to cluster analysis, and the applicability where group membership is unknown. Rough clustering solutions allow the multiple cluster membership of objects. The technique is demonstrated through the analysis of a data set containing scores associated with psychographic variables, obtained from a survey of shopping orientation and Web purchase intentions. Chapter 13 presents a survey of medical data mining focusing upon the use of heuristic techniques. The chapter proposes a forward-looking responsibility for mining practitioners that includes evaluating and justifying data mining methods–a task especially salient when heuristic methods are used. The chapter specifically considers the characteristics of medical data, reviewing a range of mining applications and approaches. In Chapter 14, machine learning techniques are used to predict the behavior of credit card users. The performance of these techniques is compared by both analyzing their correct classification rates and the knowledge extracted in a linguistic representation (rule sets or decision trees). The use of a linguistic representation for expressing knowledge acquired by learning systems aims to improve the user understanding. Under this assumption and to make sure that these systems will be accepted, several techniques have been developed by the artificial intelligence community, under both the symbolic and the connectionist approaches. The goal of Chapter 15 is to integrate evolutionary learning tools into the knowledge discovery process and to apply them to the large-scale, archaeological spatial-temporal data produced by the surveys. This heuristic approach presented in the chapter employs rough set concepts in order to represent the domain knowledge and the hypotheses.

This book will be useful to policy makers, business professionals, academics and students. We expect that the promising opportunities illustrated by the case studies and the tools and techniques described in the book will help to expand the horizons of KDD and disseminate knowledge to both the research and the practice communities.

We would like to acknowledge the help of all involved in the collation and the review process of the book, without whose support the project could not have been satisfactorily completed. Most of the authors of chapters included in this volume also served as referees for articles written by other authors. Thanks also to several other referees who have kindly refereed chapters accepted for this volume. Thanks go to all those who provided constructive and comprehensive reviews and comments. A further special note of thanks goes to all the staff at Idea Group Publishing, whose contributions throughout the whole process from inception to final publication have been invaluable.

In closing, we wish to thank all the authors for their insight and excellent contributions to this book. In addition, this book would not have been possible without the ongoing professional support from Senior Editor Dr. Mehdi Khosrowpour, Managing Editor Ms. Jan Travers and Development Editor Ms. Michele Rossi at Idea Group Publishing. Finally, we want to thank our families for their love and support throughout this project.

Ruhul Sarker, Hussein Abbass and Charles Newton
Editors