Preface

Data mining (DM) is the extraction of hidden predictive information from large databases (DBs). With the automatic discovery of knowledge implicit within DBs, DM uses sophisticated statistical analysis and modeling techniques to uncover patterns and relationships hidden in organizational DBs. Over the last 40 years, the tools and techniques to process structured information have continued to evolve from DBs to data warehousing (DW) to DM. DW applications have become business-critical. DM can extract even more value out of these huge repositories of information.

Approaches to DM are varied and often confusing. This book presents an overview of the state of art in this new and multidisciplinary field. DM is taking off for several reasons: organizations are gathering more data about their businesses, costs of storage have dropped drastically, and competitive business pressures have increased. Other factors include the emergence of pressures to control existing IT investments, and last, but not least, the marked reduction in the cost/performance ratio of computer systems. There are four basic mining operations supported by numerous mining techniques: predictive model creation supported by supervised induction techniques; link analysis supported by association discovery and sequence discovery techniques; DB segmentation supported by clustering techniques; and deviation detection supported by statistical techniques.

Although DM is still in its infancy, companies in a wide range of industries - including retail, banking and finance, heath care, manufacturing, telecommunication, and aerospace - as well as government agencies are already using DM tools and techniques to take advantage of historical data. By using pattern-recognition technologies and statistical and mathematical techniques to sift through warehoused information, DM helps analysts recognize significant facts, relationships, trends, patterns, exceptions, and anomalies that might otherwise go unnoticed.

In my February 2001 call for chapters, I sought contributions to this book that would address a vast number of issues ranging from the breakthrough of new theories to case studies of firms' experiences with their DM. After spending one and a half years of preparation on the book and a strict peer-refereed process, I am delighted to see it appearing on the market. The primary objective of this book is to explore the myriad issues regarding DM, specifically focusing on those areas that explore new methodologies or examine case studies. A broad spectrum of scientists, practitioners, graduate students, and managers, who perform research and/or implement the discoveries, are the envisioned readers of this book.

The book contains a collection of twenty chapters written by a truly international team of forty-four experts representing the leading scientists and talented young scholars from seven countries (or areas): Argentina, Canada, Italy, South Africa, Sweden, Taiwan, and the United States.

Chapter 1 by Arnborg reviews the fundamentals of inference and gives a motivation for Bayesian analysis. The method is illustrated with dependency tests in data sets with categorical data variables, and the Dirichlet prior distributions. Principles and problems for deriving causality conclusions are reviewed and illustrated with Simpson's paradox. Selection of decomposable and directed graphical models illustrates the Bayesian approach. Bayesian and Expectation Maximization (EM) classification is described briefly. The material is illustrated by two cases, one in personalization of media distribution, and one in schizophrenia research. These cases are illustrations of how to approach problems that exist in many other application areas.

Chapter 2 by Hsu discusses the problem of Feature Selection (also called Variable Elimination) in supervised inductive learning approaches to DM, in the context of controlling Inductive Bias - i.e., any preference for one (classification or regression) hypothesis other than pure consistency with training data. Feature selection can be achieved using combinatorial search and optimization approaches. This chapter focuses on data-driven validation-based techniques, particularly the WRAPPER approach. Hsu presents a wrapper that uses Genetic Algorithms for the search component and a validation criterion, based upon model accuracy and problem complexity, as the Fitness Measure. This method is related to the Classifier System of Booker, Golderberg and Holland (1989). Current research relates the Model Selection criterion in the fitness to the Minimum Description Length (MDL) family of learning criteria. Hsu presents two case studies in large-scale commercial DM and decision support: crop condition monitoring, and loss prediction for insurance pricing. Part of these case studies includes a synopsis of the general experimental framework, using the Machine Learning in Java (MLJ) and Data to Knowledge (D2K) Java-based visual programming systems for DM and information visualization.

Chapter 3 by Herna Viktor, Eric Paquet, and Gys le Roux explores the use of visual DM and virtual reality-based visualization in a cooperative learning environment. The chapter introduces a cooperative learning environment in which multiple DM tools reside and describes the ViziMine DM tool used to visualize the cooperative DM process. The aim of the ViziMine tool is twofold. Firstly, the data repository is visualized during data preprocessing and DM. Secondly, the knowledge, as obtained through DM, is assessed and modified through the interactive visualization of the cooperative DM process and its results. In this way, the user is able to assess and possibly improve the results of DM to reflect his or her domain expertise. Finally, the use of three-dimensional visualization, virtual reality-based visualization, and multimedia DM is discussed. The chapter shows how these leading-edge technologies can be used to visualize the data and its descriptors.

Feature subset selection is an important problem in knowledge discovery, not only for the insight gained from determining relevant modeling variables but also for the improved understandability, scalability, and possibly, accuracy of the resulting models. The purpose of Chapter 4 is to provide a comprehensive analysis of feature selection via evolutionary search in supervised and unsupervised learning. To achieve this purpose, Kim, Street, and Menczer first discuss a general framework for feature selection based on a new search algorithm, Evolutionary Local Selection Algorithm (ELSA). The search is formulated as a multi-objective optimization problem to examine the trade-off between the complexity of the generated solutions against their quality. ELSA considers multiple objectives efficiently while avoiding computationally expensive global comparison. The authors combine ELSA with Artificial Neural Networks (ANNs) and the EM algorithm for feature selection in supervised and unsupervised learning, respectively. Further, they show a new two-level evolutionary algorithm, Meta-Evolutionary Ensembles (MEE), in which feature selection is used to promote diversity among classifiers for ensemble classification.

Coppola and Vanneschi consider the application of parallel programming environments to develop portable and efficient high-performance DM tools. They discuss the main issues in exploiting parallelism in DM applications to improve the scalability of several mining techniques to large or geographically distributed DBs. The main focus of Chapter 5 is on parallel software engineering, showing that the skeleton-based, high-level approach can be effective both in developing portable high-performance DM kernels, and in easing their integration with other data management tools. Three test cases are described that present parallel algorithms for association rules, classification, and clustering, starting from the problem and going up to a concrete implementation. Experimental results are discussed with respect to performance and software costs. To help the integration of high-level application with existing environments, an object-oriented interface is proposed. This interface complements the parallel skeleton approach and allows the use of a number of external libraries and software modules as *external objects*, including shared-memory-distributed objects.

Rough set theory, originated by Z. Pawlak in 1982, among other applications, is a methodological tool for DM and machine learning. The main advantage of rough set theory is that it does not need any preliminary or additional information about data (such as probability distribution assumptions in probability classifier theory, grade of membership in fuzzy set theory, etc.). Numerical estimates of uncertainty of rough set theory have immediate interpretation in evidence theory (Dempster-Shafer theory). The chapter "Data Mining Based on Rough Sets" by Grzymala-Busse and Ziarko starts from fundamentals of rough set theory. Then two generalizations of rough set theory are presented: Variable Precision Rough Set Model (VPRSM) and Learning from Examples using Rough Sets (LERS). The prime concern of VPRSM is forming decision tables, while LERS produces rule sets. The two generalizations of rough set theory are independent and neither can be reduced to the other. Among many applications of LERS, those related to medical area and natural language are briefly described.

DM is based upon searching the concatenation of multiple DBs that usually contain some amount of missing data along with a variable percentage of inaccurate data, pollution, outliers, and noise. During the last four decades, statisticians have attempted to address the impact of missing data on IT. Chapter 7 by Brown and Kros commences with a background analysis, including a review of both seminal and current literature. Reasons for data inconsistency along with definitions of various types of missing data are discussed. The chapter mainly focuses on methods of addressing missing data and the impact that missing data has on the knowledge discovery process via prediction, estimation, classification, pattern recognition, and association rules. Finally, trends regarding missing data and DM are discussed, in addition to future research opportunities and concluding remarks.

In Chapter 8, Yang and Lee use a self-organizing map to cluster documents and form two feature maps. One of the map, namely the document cluster map, clusters documents according to the co-occurrence patterns of terms appeared in the documents. The other map, namely the word cluster map, is obtained by selecting the words of common interest for those documents in the same cluster. They then apply an iterative process to these maps to discover the main themes and generate hierarchies of the document clusters. The hierarchy generation and theme discovery process both utilize the synaptic weights developed after the clustering process using the self-organizing map. Thus, their technique incorporates the knowledge from the neural networks and may provide promising directions in other knowledge-discovery applications. Although this work was originally designed for text categorization tasks, the hierarchy mining process developed by these authors also poses an interesting direction in discovering and organizing unknown knowledge.

Although DM may often seem a highly effective tool for companies to be using in their business endeavors, there are a number of pitfalls and/or barriers that may impede these firms from properly budgeting for DM projects in the short term. In Chapter 9, Wang and Oppenheim indicate that the pitfalls of DM can be categorized into several distinct categories. The authors explore the issues of accessibility and usability, affordability and efficiency, scalability and adaptability, systematic patterns vs. sample-specific patterns, explanatory factors vs. random variables, segmentation vs. sampling, accuracy and cohesiveness, and standardization and verification. Finally, they present the technical challenges regarding the pitfalls of DM.

Chapter 10 by Troutt, Gribbin, Shanker, and Zhang proposes the principle of Maximum Performance Efficiency (MPE) as a contribution to the DM toolkit. This principle seeks to estimate optimal or boundary behavior, in contrast to techniques like regression analysis that predict average behavior. This MPE principle is explained and used to estimate best-practice cost rates in the context of an activity-based costing situation where the authors consider multiple activities contributing to a single cost pool. A validation approach for this estimation method is developed in terms of what the authors call normal-like-or-better performance effectiveness. Extensions to time series data on a single unit, and marginal cost-oriented basic cost models are also briefly described.

One of the major problems faced by DM technologies is how to deal with uncertainty. Bayesian methods provide an explicit way of using probability for quantifying uncertainty. The purpose of Chapter 11 by Lauria and Tayi is twofold: to provide an overview of the theoretical framework of Bayesian methods and its application to DM, with special emphasis on statistical modeling and machine learning techniques. Topics covered include Bayes Theorem and its implications, Bayesian classifiers, Bayesian belief networks, statistical computing, and an introduction to Markov Chain Monte Carlo techniques. The coverage of these topics has been augmented by providing numerical examples.

Knowledge of the structural organization of information in documents can be of significant assistance to information systems that use documents as their knowledge bases. In particular, such knowledge is of use to information retrieval systems that retrieve documents in response to user queries. Chapter 12 by Kulyukin and Burke presents an approach to mining free-text documents for structure that is qualitative in nature. It complements the statistical and machine learning approaches insomuch as the structural organization of information in documents is discovered through mining free text for content markers left behind by document writers. The ultimate objective is to find scalable DM solutions for free-text documents in exchange for modest knowledge engineering requirements.

Chapter 13 by Johnson, Fotouhi, and Draghici presents three systems that incorporate document structure information into a search of the Web. These systems extend existing Web searches by allowing the user to not only request documents containing specific search words, but also to specify that documents be of a certain type. In addition to being able to search a local DB, all three systems are capable of dynamically querying the Web. Each system applies a *query-by-structure* approach that captures and utilizes structure information as well as content during a query of the Web. Two of the systems also employ Neural Networks (NNs) to organize the information based on relevancy of both the content and structure. These systems utilize a supervised Hamming NN and an unsupervised competitive NN, respectively. Initial testing of these systems has shown promising result when compared to straight keyword searches.

Chapter 14 seeks to evaluate the feasibility of using self-organizing maps (SOMs) for financial benchmarking of companies. Eklund, Back, Vanharanta, and Visa collected a number of annual reports from companies in the international pulp and paper industry, for the period 1995-2000. They then create a financial DB consisting of a number of financial ratios, calculated based on the values from the income and balance sheets of the annual reports. The financial ratios used were selected based on their reliability and validity in international comparisons. The authors also briefly discuss issues related to the use of SOMs, such as data pre-processing, and the training of the map. The authors then perform a financial benchmarking of the companies by visualizing them on a SOM. This benchmarking includes finding the best and poorest performing companies, illustrating the effects of the Asian financial crisis, and comparing the performance of the five largest pulp and paper companies. The findings are evaluated using existing domain knowledge, i.e., information from the textual parts of the annual reports. The authors found the SOM to be a feasible tool for financial benchmarking.

In Chapter 15, general insight into DM with emphasis on the health care industry is provided by Payton. The discussion focuses on earlier electronic commerce health care initiatives, namely community health information networks (CHINs). CHINs continue to be widely debated by leading industry groups, such as The Healthy Cities Organization and The IEEE-USA Medical Technology and Policy Committee. These applications raise issues about how patient information can be mined to enable fraud detection, profitability analysis, patient profiling, and retention management. Withstanding these DM capabilities, social issues abound.

In Chapter 16, Long and Troutt discuss the potential contributions DM could make within the Human Resource (HR) function. They provide a basic introduction to DM techniques and processes and survey the literature on the steps involved in successfully mining this information. They also discuss the importance of DW and datamart considerations. A discussion of the contrast between DM and more routine statistical studies is given. They examine the value of HR information to support a firm's competitive position and for support of decision-making in organizations. Examples of potential applications are outlined in terms of data that is ordinarily captured in HR information systems. They note that few DM applications have been reported to date in the literature and hope that this chapter will spur interest among upper management and HR professionals.

The banking industry spends a large amount of IT budgets with the expectation that the investment will result in higher productivity and improved financial performance. However, bank managers make decisions on how to spend large IT budgets without accurate performance measurement systems on the business value of IT. It is a challenging DM task to investigate banking performance as a result of IT investment, because numerous financial and banking performance measures are present with the new IT cost category. Chapter 17 by Chen and Zhu presents a new DM approach that examines the impact of IT investment on banking performance, measures the financial performance of banking, and extracts performance patterns. The information obtained will provide banks with the most efficient and effective means to conduct business while meeting internal operational performance goals.

Chapter 18 by Cook and Cook highlights both the positive and negative aspects of DM. Specifically, the social, ethical, and legal implications of DM are examined through recent case law, current public opinion, and small industry-specific examples. There are many issues concerning this topic. Therefore, the purpose of this chapter is to expose the reader to some of the more interesting ones and provide insight into how information systems (ISs) professionals and businesses may protect themselves from the negative ramifications associated with improper use of data. The more experience with and exposure to social, ethical, and legal concerns with respect to DM, the better prepared the reader will be to prevent trouble in the future.

Chapter 19 by Böhm, Galli, and Chiotti presents a DM application to software engineering. Particularly, it describes the use of DM in different parts of the design process of a dynamic decision-support system agent-based architecture. By using DM techniques, a discriminating function to classify the system domains is defined. From this discriminating function, a system knowledge base is designed that stores the values of the parameters required by such a function. Also, by using DM, a data structure for analyzing the system operation results is defined. According to that, a case base to store the information of performed searches quality is designed. By mining this case base, rules to infer possible causes of domains classification error are specified. Based on these rules, a learning mechanism to update the knowledge base is designed.

DM is a field that is experiencing rapid growth and change, and new applications and developments are constantly being introduced. While many of the traditional statistical approaches to DM are still widely used, new technologies and uses for DM are coming to the forefront. The purpose of Chapter 20 is to examine and explore some of the newer areas of DM that are expected to have much impact not only for the present, but also for the future. These include the expanding areas of Web and text mining, as well as ubiquitous, distributed/collective, and phenomenal DM. From here, the discussion turns to the dynamic areas of hypertext, multimedia, spatial, and geographic DM. For those who love numbers and analytical work, constraint-based and time-eries mining are useful ways to better understand complex data. Finally, some of the most critical applications are examined, including bioinformatics.

References

Booker, L.B., Goldberg, D.E., & Holland, J.H. (1989). Classifier Systems and Genetic Algorithms. *Artificial Intelligence*, 40, 235-282.

Pawlak, Z. (1982). Rough Sets. *International Journal of Computer and Information Sciences,* 11, 341-356.