Preface

The idea of this book was conceived in June 2004, when a small group of researchers in the data-mining field gathered on the shores of the lake of Como, in Italy, to attend a focused conference–MML, *Mathematical Methods for Learning* 2004—having the objective of fostering the interaction among scholars from different countries and with different scientific backgrounds, sharing their research interests in data mining and knowledge discovery. As one of the side effects of that meeting, the conference organizers took on the exciting task of editing high quality scientific publications, where the main contributions presented at the MML conference could find an appropriate place, one next to the other, as they fruitfully did within the conference sessions. Some of the papers presented in Como, sharing a focus on mathematical optimization methods for data mining, found their place in a special issue of the international journal *Computer Optimization and Applications* (*COAP, 38*(1), 2007). Another large group of papers constituted the most appropriate building blocks for an edited book that would span a vast area of data-mining methods and applications, showing, on one hand, the relevance of mathematical methods and algorithms aimed at extracting knowledge from data, and on the other hand, how wide the application domains of data mining are. Shortly later, such project found interest and support by IGI Global, a dynamic publisher very active in promoting research-oriented publications in technological and advanced fields of knowledge. We eventually managed to finalize all the chapters, and moreover, enriched the book with additional research work that, although not presented at the MML conference, appear to have a strong relevance within the scope of the book. Most of the chapters have evolved since they were presented in 2004, and authors had the opportunity to update their work with additional results until the beginning of 2007.

The Motivations of Data Mining

The interest in data mining of researchers and practitioners with different backgrounds has increased steadily year after year. This growth is due to several reasons.

First, data mining plays today a fundamental role in analyzing and understanding the vast amount of information collected by business, government, and scientific applications. The ability to analyze large bodies of data and extract from them relevant knowledge has become a valuable service for most organizations that operate in the highly globalized and competitive business arena. The technical skills required to operate and put to use data-mining techniques are now appreciated, and often required, by the business intelligence units of financial institutions, government agencies, telecommunication companies, service providers, retailers, and distribution operators.

A second reason is to be found in the excellent and constantly improving quality of the methods and tools that are being developed in this field. Advanced mathematical models, state-of-the-art algorithmic techniques, and efficient data management systems, combined with a decreasing cost of computational power and computer memory, are now able to support data analysts with methodologies and tools that were not available a few years ago. Furthermore, such instruments are often available at low cost and with easy-to-use interfaces, integrated into well-established data management systems.

A third reason that is not to be overlooked is connected with the role that data-mining methods are playing in providing support to basic research in many scientific areas. To mention an example, biology and genetics are currently enjoying the results of the application of advanced mining techniques that allow discovery of valuable facts in complex data gathered from experiments in vitro.

Finally, we wish to mention the impulse to methodological research that has been given in many areas by the open problems posed by data-mining applications. The learning and classification problems coming from real-life problems have been exploited through many mathematical theories under different formalizations, and theoretical results of unusual relevance have been reached in optimization theory, computer science, and statistics, also thanks to the many new and stimulating problems.

Data Mining as a Practical Science

Data mining is located at the crossing of different disciplines. Its roots are to be found in the data analysis techniques that were originally the main object of the study of statistics. The fundamental ideas at the basis of estimation theory, classification, clustering, sampling theory, are indeed still one of the major ingredients of data mining. But other methods and techniques have been added to the toolbox of the data analyst, extending the limits of the classical parametric statistics with more complex models, reaching their maturity with the actual state of knowledge on decision trees, neural networks, support vector machines, just to mention a few. In addition, the need to organize and manage large bodies of data has required the deployment of computer science techniques for database management, query optimization, optimal coding of algorithms, and other tasks devoted to the storing of information in the memory of computers and to the efficient execution of algorithms.

A common trademark of the modern approaches is the formalization of estimation and classification problems arising in data mining as mathematical optimization problems, and the use of consistent algorithmic techniques to determine optimal solutions for these problems. Such methodological framework has been strongly supported by applied mathematics and operations research (OR), a scientific discipline characterized by a deep integration of mathematical theory and practical problems. A significant evidence of the role of OR in data mining is the contribution that nonlinear and integer optimization methods have given to the solution of the error minimization functions that need to be optimized to train neural networks and support vector machines. Analogously, integer programming and combinatorial optimization have been largely used to solve problems arising in the identification of synthetic rule-based classification models and in the selection of optimal subsets of features in large datasets.

Despite its strong methodological characterization, data mining cannot be successfully applied without a deep understanding of the semantic of each specific problem, which often requires the customization of existing methods or the development of ad hoc techniques, partially based on already existing algorithms. To some extent, the real challenge that the data mining practitioner has to face is the selection, among many different methods and approaches, of the one that best serves the scope of the task considered, often assessing a compromise between the complexity of the chosen model and its generalization capability.

The Contribution of this Edited Book

This book aims to provide a rich collection of current research on a broad array of topics in data mining, ranging from recent theoretical advancements in the field to relevant applications in diverse domains. Future directions and trends in data mining are also identified in most chapters.

Therefore, this volume should be an excellent guide to researchers, practitioners, and students. Its audience is represented by the research community; business executives and consultants; and senior students in the fields of data mining, information and knowledge creation, optimization, statistics, and computer science.

A Guided Tour of the Chapters

The book is composed of 19 chapters. Each one is authored by a different group of scientists, treats one of the many different theoretical or practical aspects of data mining, and is self contained with respect to the treated subject.

The first four chapters deal, to different degrees, with data-mining problems in logic setting, where the main purpose is to extract rules in logic format from the available data.

In particular, **Chapter I** is written by *Johnathan Mugan* and *Klaus Truemper*, and describes a sophisticated and complete technique to transform a set of data represented in various formats by means of an extended set of logic variables. Such task, often referred to as discretization, or binarization, is a key step in the application of logic-based classification methods to data that is described by rational or nominal variables. The chapter extends the notion of rational variables with the definition of set variables, for example, variables that are represented by their membership functions to one or more sets. The method described is characterized by the fact that the set of logic variables extracted is compact, but strongly aimed at the task of classifying, with high precision, the available data with respect to a given binary target variable. The algorithm that implements the ideas described in the chapter has been implemented and integrated into the logic data mining software Lsquare, made available by the authors as open source.

Chapter II is written by *Massimo Liguori* and *Andrea Scozzari*. Here the subject is the use of another well-known logic data mining technique, the logical analysis of data (LAD), originally developed at the University of Rutgers by the research team led by Peter Hammer. The authors propose an interesting use of this method to treat logic classification where the target variable is of ternary nature (i.e., it can assume one of three possible values). Even more interesting is the application for which the method has been developed: the financial timing decision problem, namely the problem of deciding when to buy and when to sell a given stock to maximize the profit of the trading operations. The results presented in this chapter testify how logic methods can give a significant contribution in a field where classical statistics has always played the main role.

Chapter III, authored by *Xenia Naidenova*, brings to the readers' attention several interesting theoretical aspects of logic deduction and induction that find relevant application in the construction of machine-learning algorithms. The chapter treats extensively the many details connected with this topic, and enlightens many results with simple examples. The author adopts the lattice theory as the basic mathematical tool, and succeeds in proposing a sound integration of inductive and deductive reasoning for learning implicative logic rules. The results described are the basis for the implementation of an algorithm that efficiently infers good maximally redundant tests.

In **Chapter IV**, *Giovanni Felici* and *Valerio Gatta* describe a study where the results of a stated preference model for measuring quality of service is combined with logic-based data mining to gain deeper insight in the system of preferences expressed by the customers of a large airport. The data-mining methods considered are decision trees and the logic miner Lsquare. The results are presented in the form of a set of rules that enables one to understand the similarities and the differences in two different methods to compute a quality of service index.

The topics of the following three chapters evolve around the concept of support vector machines (SVM), a mathematical method for classification and regression emerged in the last decade from statistical learning theory, which quickly attained remarkable results in many applications. SVM are based on optimization methods, particularly in the field of nonlinear programming, and are a vivid example of the contributions that can be given to data mining by state-of-the-art theoretical research in mathematical optimization.

In **Chapter V**, *Brian C. Lovell* and *Christian J. Walder* provide a rich overview of SVM in the context of data mining for business applications. They describe, with high clarity, the basic steps in SVM theory, and then integrate the chapter with several practical considerations on the use of this class of methods, comparing it with other learning approaches in the context of real-life applications.

An important role in SVM is played by kernel functions, which provide an implicit transformation of the representation of the original space of data into a high dimensional space of features. By means of such transformations, SVM can efficiently determine linear transformations in the feature space that correspond to nonlinear separations into the original space.

The identification of the right kernel function is the topic of **Chapter VI**, written by *Shawkat Ali* and *Kate A. Smith*, where they describe the application of a metalearning approach to optimally estimate the parameters that identify the kernel function before SVM is applied. The chapter highlights clearly the role of parameter

estimation in the use of learning models, and discusses how the estimation procedure should be able to adapt to the specific dataset under analysis. The experimental analysis provides tests on both binary and multicategory classification problems.

An interesting evolution of SVM is represented by discrete support vector machines, proposed in the last few years by *Carlotta Orsenigo* and *Carlo Vercellis*, authors of **Chapter VII**. According to statistical learning theory, discrete SVM directly face the minimization of the misclassification rate, within the risk functional, instead of replacing it with the misclassification distance as traditional SVM. The problem is then modeled as a mixed-integer programming problem. The method, already successful in other applications, is extended and applied here to protein folding, a very challenging task in multicategory classification. The experiments performed by the authors on benchmark datasets show that the proposed method achieves the highest accuracy in comparison to other techniques.

The use of data-mining methods to extract knowledge from large databases in genetic and biomedical applications is increasing at a fast pace, and **Chapter VIII**, written by *Li Liao*, deals with this topic. Often the data in this context is based on vectors of extremely large dimensions, and specific techniques must be deployed to obtain successful results. Li Liao tackles several of the specific problems related with handling biomedical data, in particular those related with data described by attributes that are correlated with each other and are organized in a hierarchical structure. Clustering and classification methods that exploit the hierarchies in data are considered and compared with statistical learning methods.

Chapters nine and ten both deal with clustering, a fundamental problem in nonsupervised learning. In **Chapter IX**, *Monica Chiş* discusses hierarchical clustering, where the clusters are obtained by recursively separating the data into groups of similar objects. The methods investigated belong to the family of genetic algorithms, where an initial population of chromosomes, corresponding to potential clusters, is evolved at each iteration, generating new chromosomes with the objective of minimizing a fitness function. The genetic operators adopted here are standard mutation and crossover.

Evolving on these concepts, *T. Warren Liao* presents, in **Chapter X**, a method based on genetic algorithms to cluster univariate time series. The study of time series is indeed a very central topic in data analysis, and is often overlooked in standard data-mining applications, where the main attention is addressed to multivariate data. Time series, on the other hand, present several complex aspects linked to autocorrelation and lag parameters, and surely can benefit by the use of the new methods developed in the area of data mining. Using the method of the k-medoids, the author compares the performances of three fitness functions, two distance measures, and other parameters that characterize the genetic algorithms considered. The chapter presents several experiments on data derived from cylinder-bell-funnel data and battle simulation data.

Chapter XI is a sound example of how advanced data-mining techniques can provide relevant information in production systems. *Alex Burns*, *Shital Shah*, and *Andrew Kusiak* describe the implementation of a method that integrates genetic algorithms and data mining. The results of a rule-based data-mining algorithm are evaluated and scored using a fitness function, and the related methods made available in the context of genetic algorithms. Here again we find a strong connection between data analysis and optimization techniques, and we see how certain decision problems can be successfully solved building ad hoc procedures, where methodologies and techniques from different backgrounds are deployed. The authors describe an application of the method to a power-plant boiler and highlight the contribution given to the production process.

Chapter XII is written by *Enrico Fagiuoli*, *Sara Omerino,* and *Fabio Stella*. It is an interesting work that shows how complex models derived from classical statistical techniques can play an important role in the data treatment process. The chapter describes the use of Bayesian belief networks to perform data cleaning, a relevant problem in most data-mining applications where the information available is obtained with noisy, incomplete, or error-prone procedures. Here Bayesian belief networks are used to instantiate missing values of incomplete records, to complete truncated datasets, and to detect outliers. The effectiveness of the approach is supported by numerical experiments.

Chapter XIII deals with a similar topic, data cleaning. Here, *Chuck P. Lam* and *David G. Stork* describe, in a complete and accurate way, the problem of labeling noise, requiring the identification and treatment of records

when some of the labels attached to the records are different from the correct value due to some source of noise present in the data collection process. The chapter depicts the two main problems in cleaning labeling noise: the identification of noise and the consequent revision scheme, through removal, replacement, or escalation to human supervision. In particular, the authors examine the k-nearest neighbor method to solve the identification problem, while they use probabilistic arguments to evaluate the alternative revision schemes. The public domain UCI repository is used as a source of datasets where the proposed methods are tested.

In **Chapter XIV** another tool originated in the statistical and stochastic processes environment is used to solve a relevant mining problem, clickstream analysis, which is attracting a growing attention. The problem is generated by the need to investigate the log files produced when users visit a Web site. These log files report the sequence of steps in navigation (clicks) made by the Web users. These applications can be useful for designing Web sites and for related business-oriented analysis. The authors of the chapter, *Paolo Baldini* and *Paolo Giudici*, propose to use Markov chain models to investigate the structure of the most likely navigation path in a Web site, with the objective of predicting the next step made by a Web user, based on the previous ones.

Antonino Staiano, *Lara De Vinco*, *Giuseppe Longo*, and *Roberto Tagliaferro* are the authors of **Chapter XV**. Here the topic is the visualization of complex and multidimensional data for exploration and classification purposes. The method used is based on probabilistic principal surfaces: by means of a density function in the original space, data are projected into a reduced space defined by a set of latent variables. A special case arises when the number of latent variables is equal to three and the projected space is a spherical manifold, particularly indicated to represent sparse data. Besides visualization, such reduced spaces can be used to apply classification algorithms for efficiently determining surfaces that separate groups of data belonging to different classes. Applications of the method for data in the astronomy and genetics domains are discussed.

Chapter XVI presents an unconventional application of data-mining techniques that assists spatial navigation in virtual environments. In this setting, users are able to navigate in a three-dimensional virtual space to accomplish a number of tasks. Such navigation may be difficult or inefficient for nonexperienced users, and the application discusses the use of data-mining techniques to extract knowledge from the navigation patterns of expert users, and create good navigation models. Such process is put into action by a navigation interface that implements a frequent wayfinding-sequence method. The authors, *Mehmed Kantardzic*, *Pedram Sadeghian*, and *Walaa M. Sheta*, have run experiments in simulated virtual environments, extensively discussed in the chapter.

Data-mining algorithms can be very demanding from the point of view of computational requirements, such as speed and memory, especially when large datasets are analyzed. One possible solution to deal with the computational burden is the use of parallel and distributed computing. Such an issue is the topic of **Chapter XVII**, written by *Antonio Congiusta*, *Domenico Talia*, and *Paolo Trunfio*, on the use of grid computing for distributed data mining. An integrated architecture that can properly host all the steps of the data analysis process (data management, data transfer, data mining, knowledge representation) has been designed and is presented in the chapter. The components of this data-mining-oriented middleware, termed *knowledge grid*, are described, explaining how these services can be accessed using the standard open grid architecture model.

The last two chapters of the book are devoted to two knowledge extraction methods that have received large attention in the scientific community. They extend the limits of standard machine learning theory, and can be used to build data-mining applications able to deal with unconventional applications, and to provide information in an original format. **Chapter XVIII** is about the use of fuzzy logic. Here *Nikos Pelekis*, *Babis Theodoulidis, Ioannis Kopanakis*, and *Yannis Theodoridis* cover the design of a classification heuristic scheme based on fuzzy methods. The performances of the method are analyzed by means of extensive simulated experiments. The topic of **Chapter XIX** is the method of rough sets. This method presents several noticeable features that originally characterize the rules extracted from the data. The interest of this chapter, written by *Yanbing Liu*, *Menghao Wang*, and *Jong Tang*, is also due to the application of the method to analyze and evaluate network topologies in routing problems. The application of data mining techniques in network problems associated with telecommunication problems is novel, and is likely to represent a relevant object of research in the future.