Preface

As data mining evolves into an exciting research area spanning multiple disciplines such as machine learning, artificial intelligence, bioinformatics, medicine and business intelligence, the need to apply data mining techniques to more demanding real-world problem arises. The application of data mining techniques to domains with considerable complexity has become a major hurdle for practitioners and researchers alike. The academic study of data mining typically makes assumptions such as plentiful, correctly labeled, well organized, and error free data which often do not hold. The reality is that in real-world situations, complications occur such as the data availability, data quality, data volume, data privacy and data accessibility. This presents challenges of how to apply existing data mining techniques to these new data environments from both design and implementation perspectives.

A major source of challenges is how to expand data mining algorithms into data mining systems. There is little doubt that the importance and usefulness of data mining have been well recognized by many practitioners from the outside of the data mining community, such as business administrators and medical experts. However, when referring to data mining techniques for solutions, the way people review data mining will crucially determine the success of their projects. For example, if data mining were treated just as a tool or a algorithm rather than a systematic solution, practitioners may often find their initial results unsatisfactory, partially because of the realities such as the unmanageable poor quality data, inadequate training examples or lack of integration with the domain knowledge. All these issues require that each data mining project should be customized to meet the needs of different real-world applications, hence, require a data mining practitioner to have a comprehensive understanding beyond data mining algorithms. It is expected that a review of the data mining systems on different domains will be beneficial, from both system design and implementation perspectives, to users who intend to apply data mining techniques for complete systems.

The aim of this collection is to report the results of mining real world data sets and their associated challenges in a variety of fields. When we posted the call for chapters, we were uncertain what proposals we would receive. We were happy to receive a large variety of proposals and we chose a diverse range of application areas such as software engineering, multimedia computing, biology, clinic study, finance and banking. The types of challenges each chapter addressed were a mix of the expected and unexpected. As expected submissions dealing with well known problems such as inadequate training examples and feature selection, new challenges such as mining multiple synchronized sources were explored, as well as the challenges of incorporating domain expertise in data mining process in the form of ontologies. Perhaps the most common trends mentioned in the chapters were the notion of closing the loop in the mining process such that the mining results are able to be fed back into the data set creation and an emphasis of understanding and verification of the data mining results.

Who Should Read This Book

Rather than focusing on an intensive study on data mining algorithms, the focus of this book is the real-world challenges and solutions associated with developing practical data mining systems. The contributors of the book are the data mining practitioners as well as the experts of their own domains, and what is reported here are the
techniques they actually use in their own systems. Therefore, data mining practitioners should find this book useful for assisting in the development of practical data mining applications and solving problems raised by different real-world challenges.

We believe this book can stimulate the interests of a variety of audience types such as:

- Academic research scholars with interests in data mining related issues, this book can be a reference for them to understand the realities of real-world data mining applications and motivate them to develop practical solutions.
- General data mining practitioners with focus on knowledge discovery from real-world data, this book can provide guidance on how to design a systematic solution to fulfill the goal of knowledge discovery from their data.
- General audiences or college students who want in-depth knowledge about real-world data mining applications, they may find the examples and experiences reported in the book very useful in helping them bridging the concept of data mining to real-world applications.

Organization of This Book

The entire book is divided into seven sections: Data mining in software quality modeling, knowledge discovery for genetic and medical data, data mining in mixed media data, mining image data repository, data mining and business intelligence, data mining and ontology engineering, and traditional data mining algorithms.

Section I: Data Mining in Software Quality Modeling examines the domain of software quality estimation where the availability of labeled data is severely limited. The core of the proposed study, by Seliya and Khoshgoftaar, is the NASA JP1 dataset with in excess of 10,000 software modules with the aim of predicting if a module is defective or not. Attempts to build accurate models from just the labeled data produce undesirable results, by using semi-supervised clustering and learning techniques, their techniques can improve the results significantly. At the end of the chapter, the authors also explore the interesting direction of including the user in the data mining process via interactive labeling of clusters.

Section II: Knowledge Discovery from Genetic and Medical Data consists of two contributions, which deal with applications in biology and medicine respectively. The chapter by Moore discusses the classical problems in mining biological data: feature selection and weighting. The author studies the problem of epistasis (bimolecular physical interaction) in predicting common human diseases. Two techniques are applied: multifactor dimension reduction and a filter based wrapper technique. The first approach is a classic example of feature selection, whilst the later retains all features but with a probability of being selected in the final classifier. The author then explores how to make use of the selected features to understand why some feature combinations are associated with disease and others are not. In the second chapter, Alvair, Cabrera, Caridi, and Nguyen explore applications of data-mining for pharmaceutical clinical trials, particularly for the purpose of improving clinical trial design. This is an example of what can be referred to as closed loop data mining where the data mining results must be interpretable for better trial design and so on. The authors design a decision tree algorithm that is particularly useful for the purpose of identifying the characteristics of individuals who respond considerably different than expected. The authors provide a detailed case study for analysis of the clinical trials for schizophrenia treatment.

Section III: Data Mining in Mixed Media Data focuses on the challenges of mining mixed media data. The authors, Pan, Yang, Faloutsos, and Duygulu, explore mining from various modalities (aspects) of video clips: image, audio and transcribed text. The benefit of analyzing all three sources together enables finding correlations amongst multiple sources that can be used for a variety of applications. Mixed media data present several challenges such as how to represent features and detect correlations across multiple data modalities. The authors addresses these problems by representing the data as a graph and using a random walk algorithm to find correlations. In particular, the approach requires few parameters to estimate and scales well to large datasets. The results on image captioning indicate an improvement of over 50% when compared to traditional techniques.
**Section IV: Mining Image Data Repository** discusses various issues in mining image data repositories. The chapter by Perner describes the ImageMinger, the suite of mining techniques specifically for images. A detailed case study for cell classification and in particular the identification of antinuclear autoantibodies (ANA) is also described. The chapter by Zhang, Liu, and Gruenwald discusses the applications of decision trees for remotely sensed image data in order to generate human interpretable rules that are useful for classification. The authors propose a new iterative algorithm that creates a series of linked decision trees. They verify that the algorithm is superior in interpretability and accuracy than existing techniques for Land cover data obtained from satellite images and Urban change data from southern China.

**Section V** addresses the issues of Data Mining and Business Intelligence. Data mining has a long history of being applied in financial applications, Bruckhaus and Olecka in their respective chapters describe several important challenges of the area. Bruckhaus details how to measure the fiscal impact of a predictive model by using simple metrics such as confusion tables. Several counter-intuitive insights are provided such as accuracy can be a misleading measure and an accuracy paradox due to the typical skewness in financial data. The second half of the chapter uses the introduced metrics to quantify fiscal impact. Olecka’s chapter deals with the important problem of credit scoring via modeling credit risk. Though this may appear to be a straight-forward classification or regression problem with accurate data, the author actually points out several challenges. In addition to the existing problems such as feature selection and rare event prediction, there are other domain specific issues such as multiple yet overlapping target events (bankruptcy and contractual charge-off) which are driven by different predictors. Details and modeling solutions for predicting expected dollar loss (rather than accuracy) and overcoming sample bias are reported.

**Section VI: Data Mining and Ontology Engineering** has two chapters, which are contributed by Neaga and Sidhu with collaborators respectively. This section addresses the growing area of applying data mining in areas which are not knowledge poor. In both chapters, the authors investigate how to incorporate ontologies into data mining algorithms. Neaga provides a survey of the effect of ontologies on data mining and also the effects of mining on ontologies to create a close-loop style mining process. In particular, the author attempts to answer two explicit questions: How can domain specific ontologies help in knowledge discovery and how can Web and text mining help to build ontologies. Sidhu et al. examine the problem of using protein ontologies for data mining. They begin by describing a well-known protein ontology and then describe how to incorporate this information into clustering. In particular, they show how to use ontology to create an appropriate distance matrix and consequently how to name the clusters. A case study shows the benefits of their approach in using ontologies in general.

The last section, **Section VII: Tradition Data Mining Algorithm**, deals with well known data mining problems but with atypical techniques best suited for specific applications. Hadzic and collaborators use self organized maps to perform outlier detection for applications such as noisy instance removal. Beynon handles the problem of imputing missing values and handling imperfect data by using Dempster-Shafer theory rather than traditional techniques such as expectation maximization typically used in mining. He describes his classification and ranking belief simplex (CaRBS) system and its application to replicate the bank rating schemes of organizations such as Moody’s, S&P and Fitch. A case study of how to replicate the ratings of Fitch’s individual bank rating is given. Finally, Griffiths and collaborators look at the use of rough set theory to estimating error rates by using leave-one-out, k-fold cross validation and nonparametric bootstrapping. A prototype expert system is utilized to explore the nature of each resampling technique when variable precision rough set theory (VPRS) is applied to an example data set. The software produces a series of graphs and descriptive statistics, which are used to illustrate the characteristics of each technique with regards to VPRS, and comparisons, are drawn between the results.

We hope you enjoy this collection of chapters.

*Xingquan Zhu and Ian Davidson*