Preface

The recent proliferation of affordable data gathering and storage media and powerful computing systems have provided a solid foundation for the emergence of the new field of data mining and knowledge discovery. The main goal of this fast growing field is the analysis of large, and often heterogeneous and distributed, datasets for the purpose of discovering new and potentially useful knowledge about the phenomena or systems that generated these data. Sources from which such data can come from are various natural phenomena or systems. Examples can be found in meteorology, earth sciences, astronomy, biology, social sciences, etc. On the other hand, there is another source of datasets derived mainly from business and industrial activities. This kind of data is known as "enterprise data." The common characteristic of such datasets is that the analyst wishes to analyze them for the purpose of designing a more cost-effective strategy for optimizing some type of performance measure, such as reducing production time, improving quality, eliminating wastes, and maximizing profit. Data in this category may describe different scheduling scenarios in a manufacturing environment, quality control of some process, fault diagnosis in the operation of a machine or process, risk analysis when issuing credit to applicants, management of supply chains in a manufacturing system, data for business related decision-making, just to name a few examples.

The history of data mining and knowledge discovery is only more than a decade old and its use has been spreading to various areas. It is our assertion that every aspect of an enterprise system can benefit from data mining and knowledge discovery and this book intends to show just that. It reports the recent advances in data mining and knowledge discovery of enterprise data, with focus on both algorithms and applications. The intended audience includes the practitioners who are interested in knowing more about data mining and knowledge discovery and its potential use in their enterprises, as well as the researchers who are attracted by the opportunities for methodology developments and for working with the practitioners to solve some very exciting real-world problems.

Data mining and knowledge discovery methods can be grouped into different categories depending on the type of methods and algorithms used. Thus, one may have methods that are based on artificial neural networks (ANNs), cluster analysis, decision trees, mining of association rules, tabu search, genetic algorithms (GAs), ant colony systems, Bayes networks, rule induction, etc. There are pros and cons associated with each method and it is well known that no method dominates the other methods all the time. A very critical question here is how to decide which method to choose for a particular application. We do hope that this book would provide some answers to this question.

This book is comprised of 16 chapters, written by world renowned experts in the field from a number of countries. These chapters explore the application of different methods and algorithms to different types of enterprise datasets, as depicted in Figure 1. In each chapter, various methodological and application issues which can be involved in data mining and knowledge discovery from enterprise data are discussed.

The book starts with the chapter written by Professor Liao from Louisiana State University, U.S.A., who is also one of the Editors of this book. This chapter intends to provide an extensive coverage of the work done in this field. It describes the main developments in the type of enterprise data analyzed, the mining algorithms used, and the goals of the mining analyses. The two chapters that follow the first chapter describe two important service enterprise applications, i.e., credit rating and detection of insolvent customers. The following eight chapters deal with the mining of various manufacturing enterprise data. These application chapters are arranged in the order of activities carried out by each functional area of a manufacturing enterprise in order to fulfill customers' orders; that is, sales forecasting, process engineering, production control, process monitoring and control, fault diagnosis, quality improvement, and maintenance. Each covered area is important in its own way to the successful operation of an enterprise. The next two chapters address two unique data: one on workflow and the other on images of cell-based assays. The remaining three chapters focus more on the methodology and methodological issues. A more detailed overview of each chapter follows.

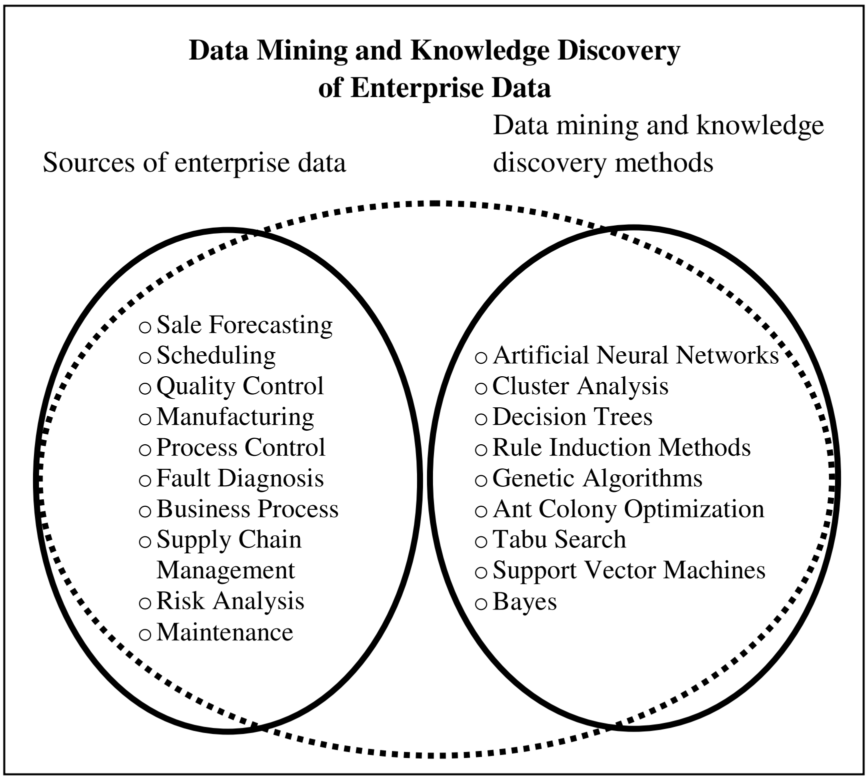

Figure 1. A sketch of data mining and knowledge discovery of enterprise data.

In particular, the second chapter is written by Professors Yu and Chen and their associates from Tsinghua University in Beijing, China. It studies some key classification methods, including decision trees, Bayesian networks, support vector machines, neural networks, *k*-nearest neighbors, and an associative classification method in analyzing credit risk of companies. A comparative study on a real dataset on credit risk reveals that the proposed associative classification method consistently outperformed all the others.

The third chapter is authored by Professors Daskalaki and Avouris from University of Patras, Greece, along with their collaborator, Mr. Kopanas. It discusses various aspects of the data mining and knowledge discovery process, particularly on imbalanced class data and cost-based evaluation, in mining customer behavior patterns from customer data and their call records.

The fourth chapter is written by Professors Chang and Wang from Yuan-Ze University and Ching-Yun University in Taiwan, respectively. In this chapter, the authors study the use of gray relation analysis for selecting time series variables and several methods, including Winter's method, multiple regression analysis, back propagation neural networks, evolving neural networks, evolving fuzzy neural networks, and weighted evolving fuzzy neural networks, for sale forecasting.

The fifth chapter is contributed by Professor Jiao and his associates from the Nanyang Technological University, Singapore. It describes how to apply specific data mining techniques such as text mining, tree matching, fuzzy clustering, and tree unification on the process platform formation problem in order to produce a variety of customized products.

The sixth chapter is written by Dr. Min and Professor Yih from Sandia National Labs and Purdue University in the U.S.A., respectively. This chapter describes a data mining approach to obtain a dispatching strategy for a scheduler so that the appropriate dispatching rules can be selected for different situations in a complex semiconductor wafer fabrication system. The methods used are based on simulation and competitive neural networks.

The seventh chapter is contributed by Professor Last and his associates from Ben-Gurion University of the Negev, Israel. It describes their application of single-objective and multi-objective classification algorithms for the prediction of grape and wine quality in a multi-year agricultural database maintained by Yarden – Golan Heights Winery in

Katzrin, Israel. This chapter indicates the potential of some data mining techniques in such diverse domains as in agriculture.

The eighth chapter is written by Professor Chien and his associates from the National Tsing Hua University, Taiwan. This chapter aims at describing characteristics of various data mining empirical studies in semiconductor manufacturing, particularly defect diagnosis and yield enhancement, from engineering data and manufacturing data.

The ninth chapter is contributed by Professors Porzio and Ragozini from the University of Cassino and the University of Naples in Italy, respectively. This chapter aims at presenting their data mining vision on Statistical Process Control (SPC) analysis and to describe their nonparametric multivariate control scheme based on the data depth approach.

The tenth chapter is written by Professor Jeong and his associates from the University of Tennessee in the U.S.A. This chapter addresses the problems of fault diagnosis based on the analysis of multi-dimensional function data such as time series and hyperspectral images. It presents some wavelet-based data reduction procedures that balance the reconstruction error against the reduction efficiency. It evaluates the performance of two approaches: partial least squares and principal component regression for shaft alignment prediction. In addition, it describes an analysis of hyperspectral images for the detection of poultry skin tumors, focusing in particular on data reduction using PCA and 2D wavelet analysis and support vector machines based classification.

In the eleventh chapter, Professor Khoo and his associates from the Nanyang Technological University in Singapore describe a hybrid approach that is based on rough sets, tabu search and genetic algorithms. The applicability of this hybrid approach is demonstrated with a case study on the maintenance of heavy machinery. The proposed hybrid approach is shown to be more powerful than the component methods when they are applied alone.

The twelfth chapter describes some recently proposed techniques of high potential for optimizing business processes and their corresponding workflow models by analyzing the details of previously executed processes, stored as a workflow log. This chapter is authored by

Professor Gunopulos from the University of California at Riverside and his collaborator from Google Inc. in the U.S.A.

The thirteenth chapter, contributed by Dr. Perner from the Institute of Computer Vision and Applied Computer Science in Germany, presents some intriguing new intelligent and automatic image analysis and interpretation procedures and demonstrates them in the application of HEp-2 cell pattern analysis, based on their *Cell_Interpret* system. Although bio-image data are mined in this chapter, the described system can be extended to other types of images encountered in other enterprise systems.

The fourteenth chapter is written by Professor Trafalis and his research associate from the University of Oklahoma in the U.S.A. The main focus of this chapter is the theoretical study of support vector machines (SVMs). These optimization methods are in the interface of operations research (O.R.) and artificial intelligence methods and seem to possess great potential. The same chapter also discusses some application issues of SVMs in sciences, business and engineering.

The fifteenth chapter, written by Professor Huo and his associates from Georgia Tech in the U.S.A., discusses some manifold-based learning methods such as local linear embedded (LLE), ISOMAP, Laplacian Eigenmaps, Hessian Eigenmap, and Local Tangent Space Alignment (LTSA), along with some important applications. These methods are relatively new compared to other methods and their potential for enterprise data mining is thus relatively unexplored.

The sixteenth chapter describes mining methods that are based on some statistical approaches. It is written by Professor Feng and his research associate from the Bradley University in the U.S.A. The statistical methods studied in this chapter include regression analysis, bootstrap, bagging, and clustering. It is shown how these methods could be used together to build an accurate model when only small datasets are available. This chapter is thus particularly relevant when there is a lack of data due to high cost or other reasons.

Each chapter is self-contained and addresses an important issue that is related to data mining methods and the analysis of enterprise data. Each chapter provides a comprehensive treatment of the topic it covers.

Furthermore, when all the chapters are considered together, they cover all aspects of crucial importance to any modern enterprise in today's increasingly competitive world.

This book is unique in that it focuses on the key algorithmic and application issues in the mining of enterprise data. Instead of discussing a particular software environment, which may become obsolete when the new version becomes available, it studies the fundamental issues related to the mining of enterprise data. A few chapters present new methodologies that are not even available in commercially available software packages at all. Thus, this book can definitely be very valuable to researchers and practitioners in the field. It can also be used by graduate students in computer science, business, or engineering schools as well.

T. Warren Liao, Ph.D.
Evangelos Triantaphyllou, Ph.D.
Louisiana State University
Baton Rouge, LA, U.S.A.

July of 2007