Preface

The IEEE ICDM 2004 workshop on the Foundation of Data Mining and the IEEE ICDM 2005 workshop on the Foundation of Semantic Oriented Data and Web Mining focused on topics ranging from the foundations of data mining to new data mining paradigms. The workshops brought together both data mining researchers and practitioners to discuss these two topics while seeking solutions to long standing data mining problems and stimulating new data mining research directions. We feel that the papers presented at these workshops may encourage the study of data mining as a scientific field and spark new communications and collaborations between researchers and practitioners.

To express the visions forged in the workshops to a wide range of data mining researchers and practitioners and foster active participation in the study of foundations of data mining, we edited this volume by involving extended and updated versions of selected papers presented at those workshops as well as some other relevant contributions. The content of this book includes studies of foundations of data mining from theoretical, practical, algorithmical, and managerial perspectives. The following is a brief summary of the papers contained in this book.

The first paper "Compact Representations of Sequential Classification Rules," by Elena Baralis, Silvia Chiusano, Riccardo Dutto, and Luigi Mantellini, proposes two compact representations to encode the knowledge available in a sequential classification rule set by extending the concept of closed itemset and generator itemset to the context of sequential rules. The first type of compact representation is called classification rule cover (CRC), which is defined by the means of the concept of generator sequence and is equivalent to the complete rule set for classification purpose. The second type of compact representation, which is called compact classification rule set (CCRS), contains compact rules characterized by a more complex structure based on closed sequence and their associated generator sequences. The entire set of frequent sequential classification rules can be re-generated from the compact classification rules set.

A new subspace clustering algorithm for high dimensional binary valued dataset is proposed in the paper "An Algorithm for Mining Weighted Dense Maximal 1-Complete Regions" by Haiyun Bian and Raj Bhatnagar. To discover patterns in all subspace including sparse ones, a weighted density measure is used by the algorithm to adjust density thresholds for clusters according to different density values of different subspaces. The proposed clustering algorithm is able to find all patterns satisfying a minimum weighted density threshold in all subspaces in a time and memory efficient way. Although presented in the context of the subspace clustering problem, the algorithm can be applied to other closed set mining problems such as frequent closed itemsets and maximal biclique.

In the paper "Mining Linguistic Trends from Time Series" by Chun-Hao Chen, Tzung-Pei Hong, and Vincent S. Tseng, a mining algorithm dedicated to extract human understandable linguistic trend from time series is proposed. This algorithm first transforms data series to an angular series based on angles of adjacent points in the time series. Then predefined linguistic concepts are used to fuzzify each angle value. Finally, the Aprori-like fuzzy mining algorithm is used to extract linguistic trends.

In the paper "Latent Semantic Space for Web Clustering" by I-Jen Chiang, T.Y. Lin, Hsiang-Chun Tsai, Jau-Min Wong, and Xiaohua Hu, latent semantic space in the form of some geometric structure in combinatorial topology and hypergraph view, has been proposed for unstructured document clustering. Their clustering work is based on a novel view that term associations of a given collection of documents form a simplicial complex, which can be decomposed into connected components at various levels. An agglomerative method for finding geometric maximal connected components for document clustering is proposed. Experimental results show that the proposed method can effectively solve polysemy and term dependency problems in the field of information retrieval.

The paper "A Logical Framework for Template Creation and Information Extraction" by David Corney, Emma Byrne, Bernard Buxton, and David Jones proposes a theoretical framework for information extraction, which allows different information extraction systems to be described, compared, and developed. This framework develops a formal characterization of templates, which are textual patterns used to identify information of interest, and proposes approaches based on AI search algorithms to create and optimize templates in an automated way. Demonstration of a successful implementation of the proposed framework and its application on biological information extraction are also presented as a proof of concepts.

Both probability theory and Zadeh fuzzy system have been proposed by various researchers as foundations for data mining. The paper "A Probability Theory Perspective on the Zadeh Fuzzy System" by Q.S. Gao, X.Y. Gao, and L. Xu conducts a detailed analysis on these two theories to reveal their relationship. The authors prove that the probability theory and Zadeh fuzzy system perform equivalently in computer reasoning that does not involve

complement operation. They also present a deep analysis on where the fuzzy system works and fails. Finally, the paper points out that the controversy on "complement" concept can be avoided by either following the additive principle or renaming the complement set as the conjugate set.

In the paper "Three Approaches to Missing Attribute Values: A Rough Set Perspective" by Jerzy W. Grzymala-Busse, three approaches to missing attribute values are studied using rough set methodology, including attribute-value blocks, characteristic sets, and characteristic relations. It is shown that the entire data mining process, from computing characteristic relations through rule induction, can be implemented based on attribute-value blocks. Furthermore, attribute-value blocks can be combined with different strategies to handle missing attribute values.

The paper "MLEM2 Rule Induction Algorithms: With and Without Merging Intervals" by Jerzy W. Grzymala-Busse compares the performance of three versions of the learning from example module of a data mining system called LERS (learning from examples based on rough sets) for rule induction from numerical data. The experimental results show that the newly introduced version, MLEM2 with merging intervals, produces the smallest total number of conditions in rule sets.

To overcome several common pitfalls in a business intelligence project, the paper "Towards a Methodology for Data Mining Project Development: the Importance of Abstraction" by P. González-Aranda, E. Menasalves, S. Millán, Carlos Ruiz, and J. Segovia proposes a data mining lifecycle as the basis for proper data mining project management. Concentration is put on the project conception phase of the lifecycle for determining a feasible project plan.

The paper "Finding Active Membership Functions in Fuzzy Data Mining" by Tzung-Pei Hong, Chun-Hao Chen, Yu-Lung Wu, and Vincent S. Tseng proposes a novel GA-based fuzzy data mining algorithm to dynamically determine fuzzy membership functions for each item and extract linguistic association rules from quantitative transaction data. The fitness of each set of membership functions from an itemset is evaluated by both the fuzzy supports of the linguistic terms in the large 1-itemsets and the suitability of the derived membership functions, including overlap, coverage, and usage factors.

Improving the efficiency of mining frequent patterns from very large datasets is an important research topic in data mining. The way in which the dataset and intermediary results are represented and stored plays a crucial role in both time and space efficiency. The paper "A Compressed Vertical Binary Algorithm for Mining Frequent Patterns" by J. Hdez. Palancar, R. Hdez. León, J. Medina Pagola, and A. Hechavarría proposes a compressed vertical binary representation of the dataset and presents approach to mine frequent patterns based on this representation. Experimental results show that the compressed vertical binary approach outperforms Apriori, optimized Apriori, and Mafia on several typical test datasets.

Causal reasoning plays a significant role in decision-making, both formally and informally. However, in many cases, knowledge of at least some causal

effects is inherently inexact and imprecise. The chapter "Naïve Rules Do Not Consider Underlying Causality" by Lawrence J. Mazlack argues that it is important to understand when association rules have causal foundations in order to avoid naïve decisions and increases the perceived utility of rules with causal underpinnings. In his second chapter "Inexact Multiple-Grained Causal Complexes", the author further suggests using nested granularity to describe causal complexes and applying rough sets and/or fuzzy sets to soften the need for preciseness. Various aspects of causality are discussed in these two chapters.

Seeing the needs for more fruitful exchanges between data mining practice and data mining research, the paper "Does Relevance Matter to Data Mining Research" by Mykola Pechenizkiy, Seppo Puuronen, and Alexcy Tsymbal addresses the balance issue between the rigor and relevance constituents of data mining research. The authors suggest the study of the foundation of data mining within a new proposed research framework that is similar to the ones applied in the IS discipline, which emphasizes the knowledge transfer from practice to research.

The ability to discover actionable knowledge is a significant topic in the field of data mining. The paper "E-Action Rules" by Li-Shiang Tsay and Zbigniew W. Ras proposes a new class of rules called "E-action rules" to enhance the traditional action rules by introducing its supporting class of objects in a more accurate way. Compared with traditional action rules or extended action rules, e-action rule is easier to interpret, understand, and apply by users. In their second paper "Mining e-Action Rules, System DEAR," a new algorithm for generating e-action rules, called Action-tree algorithm is presented in detail. The action tree algorithm, which is implemented in the system DEAR2.2, is simpler and more efficient than the action-forest algorithm presented in the previous paper.

In his first paper "Definability of Association Rules and Tables of Critical Frequencies," Jan Ranch presents a new intuitive criterion of definability of association rules based on tables of critical frequencies, which are introduced as a tool for avoiding complex computation related to the association rules corresponding to statistical hypotheses tests. In his second paper "Classes of Association Rules: An Overview," the author provides an overview of important classes of association rules and their properties, including logical aspects of calculi of association rules, evaluation of association rules in data with missing information, and association rules corresponding to statistical hypotheses tests.

In the paper "Knowledge Extraction from Microarray Datasets Using Combined Multiple Models to Predict Leukemia Types" by Gregor Stiglic, Nawaz Khan, and Peter Kokol, a new algorithm for feature extraction and classification on microarray datasets with the combination of the high accuracy of ensemble-based algorithms and the comprehensibility of a single decision tree is proposed. Experimental results show that this algorithm is able

to extract rules by describing gene expression differences among significantly expressed genes in leukemia.

In the paper "Using Association Rules for Classification from Databases Having Class Label Ambiguities: A Belief Theoretic Method" by S.P. Subasinghua, J. Zhang, K. Premaratae, M.L. Shyu, M. Kubat, and K.K.R.G.K. Hewawasam, a classification algorithm that combines belief theoretic technique and portioned association mining strategy is proposed, to address both the presence of class label ambiguities and unbalanced distribution of classes in the training data. Experimental results show that the proposed approach obtains better accuracy and efficiency when the above situations exist in the training data. The proposed classifier would be very useful in security monitoring and threat classification environments where conflicting expert opinions about the threat category are common and only a few training data instances available for a heightened threat category.

Privacy preserving data mining has received ever-increasing attention during the recent years. The paper "On the Complexity of the Privacy Problem" explores the foundations of the privacy problem in databases. With the ultimate goal to obtain a complete characterization of the privacy problem, this paper develops a theory of the privacy problem based on recursive functions and computability theory.

In the paper "Ensembles of Least Squares Classifiers with Randomized Kernels," the authors, Kari Torkkola and Eugene Tuv, demonstrate that stochastic ensembles of simple least square classifiers with randomized kernel widths and OOB-past-processing achieved at least the same accuracy as the best single RLSC or an ensemble of LSCs with fixed tuned kernel width, but require no parameter tuning. The proposed approach to create ensembles utilizes fast exploratory random forests for variable filtering as a preprocessing step; therefore, it can process various types of data even with missing values.

Shusahu Tsumoto contributes two papers that study contigency table from the perspective of information granularity. In the first paper "On Pseudo-statistical Independence in a Contingency Table," Shusuhu shows that a contingency table may be composed of statistical independent and dependent parts and its rank and the structure of linear dependence as Diophatine equations play very important roles in determining the nature of the table. The second paper "Role of Sample Size and Determinants in Granularity of Contingency Matrix" examines the nature of the dependence of a contingency matrix and the statistical nature of the determinant. The author shows that as the sample size N of a contingency table increases, the number of 2×2 matrix with statistical dependence will increase with the order of N^3, and the average of absolute value of the determinant will increase with the order of N^2.

The paper "Generating Concept Hierarchy from User Queries" by Bob Wall, Neal Richter, and Rafal Angryk develops a mechanism that builds concept hierarchy from phrases used in historical queries to facilitate users' navigation of the repository. First, a feature vector of each selected query is generated by extracting phrases from the repository documents matching the

query. Then the Hierarchical Agglomarative Clustering algorithm and subsequent portioning and feature selection and reduction processes are applied to generate a natural representation of the hierarchy of concepts inherent in the system. Although the proposed mechanism is applied to an FAQ system as proof of concept, it can be easily extended to any IR system.

Classification Association Rule Mining (CARM) is the technique that utilizes association mining to derive classification rules. A typical problem with CARM is the overwhelming number of classification association rules that may be generated. The paper "Mining Efficiently Significant Classification Associate Rules" by Yanbo J. Wang, Qin Xin, and Frans Coenen addresses the issues of how to efficiently identify significant classification association rules for each predefined class. Both theoretical and experimental results show that the proposed rule mining approach, which is based on a novel rule scoring and ranking strategy, is able to identify significant classification association rules in a time efficient manner.

Data mining is widely accepted as a process of information generalization. Nevertheless, the questions like what in fact is a generalization and how one kind of generalization differs from another remain open. In the paper "Data Preprocessing and Data Mining as Generalization" by Anita Wasilewska and Ernestina Menasalvas, an abstract generalization framework in which data preprocessing and data mining proper stages are formalized as two specific types of generalization is proposed. By using this framework, the authors show that only three data mining operators are needed to express all data mining algorithms; and the generalization that occurs in the preprocessing stage is different from the generalization inherent to the data mining proper stage.

Unbounded, ever-evolving and high-dimensional data streams, which are generated by various sources such as scientific experiments, real-time production systems, e-transactions, sensor networks, and online equipments, add further layers of complexity to the already challenging "drown in data, starving for knowledge" problem. To tackle this challenge, the paper "Capturing Concepts and Detecting Concept-Drift from Potential Unbounded, Ever-Evolving and High-Dimensional Data Streams" by Ying Xie, Ajay Ravichandran, Hisham Haddad, and Katukuri Jayasimha proposes a novel integrated architecture that encapsulates a suit of interrelated data structures and algorithms which support (1) real-time capturing and compressing dynamics of stream data into space-efficient synopses and (2) online mining and visualizing both dynamics and historical snapshots of multiple types of patterns from stored synopses. The proposed work lays a foundation for building a data stream warehousing system as a comprehensive platform for discovering and retrieving knowledge from ever-evolving data streams.

In the paper "A Conceptual Framework of Data Mining," the authors, Yiyu Yao, Ning Zhong, and Yan Zhao emphasize the need for studying the nature of data mining as a scientific field. Based on Chen's three-dimension view, a threelayered conceptual framework of data mining, consisting of the philosophy layer, the technique layer, and the application layer, is discussed

in their paper. The layered framework focuses on the data mining questions and issues at different abstract levels with the aim of understanding data mining as a field of study, instead of a collection of theories, algorithms, and software tools.

The papers "How to Prevent Private Data from Being Disclosed to a Malicious Attacker" and "Privacy-Preserving Naive Bayesian Classification over Horizontally Partitioned Data" by Justin Zhan, LiWu Chang, and Stan Matwin, address the issue of privacy preserved collaborative data mining. In these two papers, secure collaborative protocols based on the semantically secure homomorphic encryption scheme are developed for both learning Support Vector Machines and Nave Bayesian Classifier on horizontally partitioned private data. Analyses of both correctness and complexity of these two protocols are also given in these papers.

We thank all the contributors for their excellent work. We are also grateful to all the referees for their efforts in reviewing the papers and providing valuable comments and suggestions to the authors. It is our desire that this book will benefit both researchers and practitioners in the filed of data mining.

Tsau Young Lin
Ying Xie
Anita Wasilewska
Churn-Jung Liau

http://www.springer.com/978-3-540-78487-6

Data Mining: Foundations and Practice

Lin, T.Y.; Xie, Y.; Wasilewska, A.; Liau, C.-J. (Eds.)

2008, XVI, 562 p. 129 illus., 25 illus. in color., Hardcover
ISBN: 978-3-540-78487-6