

---

## Preface

The growth in the amount of data collected and generated has exploded in recent times with the widespread automation of various day-to-day activities, advances in high-level scientific and engineering research and the development of efficient data collection tools. This has given rise to the need for automatically analyzing the data in order to extract knowledge from it, thereby making the data potentially more useful.

Knowledge discovery and data mining (KDD) is the process of identifying valid, novel, potentially useful and ultimately understandable patterns from massive data repositories. It is a multi-disciplinary topic, drawing from several fields including expert systems, machine learning, intelligent databases, knowledge acquisition, case-based reasoning, pattern recognition and statistics.

Many data mining systems have typically evolved around well-organized database systems (e.g., relational databases) containing relevant information. But, more and more, one finds relevant information hidden in unstructured text and in other complex forms. Mining in the domains of the world-wide web, bioinformatics, geoscientific data, and spatial and temporal applications comprise some illustrative examples in this regard. Discovery of knowledge, or potentially useful patterns, from such complex data often requires the application of advanced techniques that are better able to exploit the nature and representation of the data. Such advanced methods include, among others, graph-based and tree-based approaches to relational learning, sequence mining, link-based classification, Bayesian networks, hidden Markov models, neural networks, kernel-based methods, evolutionary algorithms, rough sets and fuzzy logic, and hybrid systems. Many of these methods are developed in the following chapters.

In this book, we bring together research articles by active practitioners reporting recent advances in the field of knowledge discovery, where the information is mined from complex data, such as unstructured text from the world-wide web, databases naturally represented as graphs and trees, geoscientific data from satellites and visual images, multimedia data and bioinformatic data. Characteristics of the methods and algorithms reported here include the use of domain-specific knowledge for reducing the search space, dealing with

uncertainty, imprecision and concept drift, efficient linear and/or sub-linear scalability, incremental approaches to knowledge discovery, and increased level and intelligence of interactivity with human experts and decision makers. The techniques can be sequential, parallel or stream-based in nature.

The book has been divided into two main sections: foundations and applications. The chapters in the foundations section present general methods for mining complex data. In Chapter 1, Bandyopadhyay and Maulik present an overview of the field of data mining and knowledge discovery. They discuss the main concepts of the field, the issues and challenges, and recent trends in data mining, which provide the context for the subsequent chapters on methods and applications.

In Chapter 2, Ghosh, Kumar and Crawford address the issue of high dimensionality in both the attributes and class values of complex data. Their approach builds a binary hierarchical classifier by decomposing the set of classes into smaller partitions and performing a two-class learning problem between each partition. The simpler two-class learning problem often allows a reduction in the dimensionality of the attribute space. Their approach shows improvement over other approaches to the multi-class learning problem and also results in the discovery of knowledge in the form of the class hierarchy.

Cook, Holder, Coble and Potts describe techniques for mining complex data represented as a graph in Chapter 3. Many forms of complex data involve entities, their attributes, and their relationships to other entities. It is these relationships that make appropriate a graph representation of the data. The chapter describes numerous techniques based on the core Subdue methodology that uses data compression as a metric for interestingness in mining knowledge from the graph data. These techniques include supervised and unsupervised learning, clustering and graph grammar learning. They address efficiency issues by introducing an incremental approach to processing streaming graph data. They also introduce a method for mining graphs in which relevant examples are embedded, possibly overlapping, in one large graph. Numerous successes are documented in a number of domains.

In Chapter 4, Gärtner also presents techniques for mining graph data, but these techniques are based on kernel methods which implicitly map the graph data to a higher-dimensional, non-relational space where learning is easier, thus avoiding the computational complexity of graph operations for matching and covering. While kernel methods have been applied to single graphs, Gärtner introduces kernels that apply to sets of graphs and shows their effectiveness on problems from the fields of relational reinforcement learning and molecular classification.

While graphs represent one of the most expressive forms of complex data representations, some specializations of graphs (e.g., trees) still allow the representation of significant relational information, but with reduced computational cost. In Chapter 5, Zaki presents a technique called TREEMINER for finding all frequent subtrees in a forest of trees and compares this approach to a pattern-matching approach. Zaki shows results indicating a significant

increase in speed over the pattern-matching approach and applies the new technique to the problem of mining usage patterns from real logs of website browsing behavior.

Another specialized form in which complex data might be expressed is a sequence. In Chapter 6, Sarawagi discusses several methods for mining sequence data, i.e., data modeled as a sequence of discrete multi-attribute records. She reviews state-of-the-art techniques in sequence mining and applies these to two real applications: address cleaning and information extraction from websites.

In Chapter 7, Getoor returns to the more general graph representation of complex data, but includes probabilistic information about the distribution of links (or relationships) between entities. Getoor uses a structured logistic regression model to learn patterns based on both links and entity attributes. Results in the domains of web browsing and citation collections indicate that the use of link distribution information improves classification performance.

The remaining chapters constitute the applications section of the book. Significant successes have been achieved in a wide variety of domains, indicating the potential benefits of mining complex data, rather than applying simpler methods on simpler transformations of the data. Chapter 8 begins with a contribution by Zhang and Wang describing techniques for mining evolutionary trees, that is, trees whose parent-child relationships represent actual evolutionary relationships in the domain of interest. A good example, and one to which they apply their approach, is phylogenetic trees that describe the evolutionary pathways of species at the molecular level. Their algorithm efficiently discovers “cousin pairs,” which are two nodes sharing a common ancestor, in a single tree or a set of trees. They present numerous experimental results showing the efficiency and effectiveness of their approach in both synthetic and real domains, namely, phylogenetic trees.

In Chapter 9, Jiang and Tan apply a variant of the *A priori*-based association rule-mining algorithm to the relational domain of Resource Description Framework (RDF) documents. Their approach treats RDF relations as items in the traditional association-rule mining framework. Their approach also takes advantage of domain ontologies to provide generalizations of the RDF relations. They apply their technique to a synthetically-generated collection of RDF documents pertaining to terrorism and show that the method discovers a small set of association rules capturing the main associations known to be present in the domain.

Saha, Das and Chanda address the task of content-based image retrieval by mapping image data into complex data using features based on shape, texture and color in Chapter 10. They also develop an image retrieval similarity measure based on human perception and improve retrieval accuracy using feedback to establish the relevance of the various features. The authors empirically validate the superiority of their method over competing methods of content-based image retrieval using two large image databases.

In Chapter 11, Mukkamala and Sung turn to the problem of intrusion detection. They perform a comparative analysis of three advanced mining

methods: support vector machines, multivariate adaptive regression splines, and linear genetic programs. Overall, they found that the three methods performed similarly on the intrusion detection problem. However, they also found that a significant increase in performance was possible using feature selection, where the above three mining methods were used to rank features by relevance. Their conclusions are empirically validated using the DARPA intrusion detection benchmark database.

One scenario affecting the above methods for mining complex data is the increasing likelihood that data will be collected via a continuous stream. In Chapter 12, Gaber, Krishnaswamy and Zaslavsky present a theoretical framework for mining algorithms applied to this scenario based on a model of on-board, resource-constrained mining. They apply their model to the task of on-board mining of data streams in sensor networks. In addition to this general framework they have also developed lightweight mining algorithms for clustering, classification and frequent itemset discovery. Their model and algorithms are empirically validated using synthetic streaming data and the resource-constrained environment of a common handheld computer.

Finally, in Chapter 13, Yang, Yan, Han and Wang also consider the task of mining data streams. They specifically focus on the constraints that the mining algorithm scan the data only once and adapt to evolving patterns present in the data stream. They develop an evolutionary classifier based on a naive Bayesian classifier and employ a train-and-test method combined with a divergence measure to detect evolving characteristics of the data stream. They perform extensive empirical testing based on synthetic data to show the efficiency and effectiveness of their approach.

In summary, the chapters on the foundations and applications of mining complex data provide a representative selection of the available methods and their evaluation in real domains. While the field is rapidly evolving into new algorithms and new types of complex data, these chapters clearly indicate the importance and potential benefit of developing such algorithms to mine complex data. The book may be used either in a graduate level course as part of the subject of data mining, or as a reference book for research workers working in different aspects of mining complex data.

We take this opportunity to thank all the authors for contributing chapters related to their current research work that provide the state of the art in advanced methods for mining complex data. We are grateful to Mr S. Santra of Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India, for providing technical assistance during the preparation of the final manuscript. Finally, a vote of thanks to Ms Catherine Drury of Springer Verlag London Ltd. for her initiative and constant support.

January, 2005

*Sanghamitra Bandyopadhyay*  
*Ujjwal Maulik*  
*Lawrence B. Holder*  
*Diane J. Cook*

Advanced Methods for Knowledge Discovery from  
Complex Data

Maulik, U.; Holder, L.B.; Cook, D.J. (Eds.)

2005, XVIII, 369 p., Hardcover

ISBN: 978-1-85233-989-0