Preface

Data Mining and Knowledge Discovery in databases have been attracting a significant amount of research, industry, and media attention of late. Data Mining may be defined as the process of extracting trends or patterns from data and the technique. Data Mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large datasets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

Knowledge Discovery (KD) may be characterized as the process that can be applied to the results of data mining, to make sense of them. A KD process includes data warehousing, target data selection, cleaning, preprocessing, transformation and reduction, data mining, model selection, evaluation and interpretation, and finally consolidation and use of the extracted knowledge. Specifically, data mining aims to develop algorithms for extracting new patterns from the facts recorded in a database. Hitherto, data mining tools adopted techniques from statistics, neural network modeling, and visualization to classify data and identify patterns. Ultimately, KD aims to enable an information system to transform information to knowledge through hypothesis testing and theory information. It sets new challenges for database technology: new concepts and methods are needed for basic operations, query languages, and query processing strategies.

Recent progress in scientific and engineering applications has accumulated huge volumes of high-dimensional data, stream data, and spatial and temporal data. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of Knowledge Discovery in Databases (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Subsequently, the KDD process can be viewed as a multidisciplinary activity that encompasses techniques beyond the scope of any one particular discipline such as machine learning. KDD places a special emphasis on finding understandable, patterns that can be interpreted as useful or interesting knowledge. KDD also emphasizes scaling and robustness properties of modeling algorithms for large noisy data sets. Related AI research fields include machine discovery, which targets the discovery of empirical laws from observation and experimentation (Shrager & Langley, 1990), glossary of terms common to KDD and machine discovery (Kloesgen & Zytkow, 1996), and causal modeling for the inference of causal models from data (Spirtes, Glymour, & Scheines, 1993). Statistics in particular has much in common with KDD (Glymour, Madigan, Pregibon, & Smyth, 1996).

A number of advances in technology and business processes have contributed to a growing interest in data mining in both the public and private sectors. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related
techniques such as neural networks and advanced algorithms; the spread of the client/server computing model, allowing users to access centralized data resources from the desktop; and an increased ability to combine data from disparate sources into a single searchable source (Makulowich, 1999).

In addition to these improved data management tools, the increased availability of information and the decreasing costs of storing it have also played a role. Over the past several years there has been a rapid increase in the volume of information collected and stored, with some observers suggesting that the quantity of the world’s data approximately doubles every year. At the same time, the costs of data storage have decreased significantly from dollars per megabyte to pennies per megabyte. Similarly, computing power has continued to double every 18-24 months, while the relative cost of computing power has continued to decrease.

KDD is an attempt to address a problem that the digital information era made a fact of life for all of us: data overload and various KDD applications have been deployed in operational use on large-scale real-world problems in science and in business. In science, main KDD application areas includes, astronomy, biomedical engineering, telecommunications, geospatial data and climate data and the earth ecosystems. In business, main KDD application areas includes marketing, finance, fraud detection, manufacturing, telecommunication, and Internet agents. These are just a few of the numerous such systems that use KDD techniques to automatically produce useful information from large mass of raw data.

Along these approaches to enable practitioners in improving their researches and participate actively in solving practical problems related to various knowledge practices and emerging applications of data mining a complete reference will be an essential need. A book featuring all these aspects can fill an extremely demanding knowledge gap in the contemporary world.

Furthermore, in selecting potential KDD applications, which can be divided into practical and technical categories various criteria should be followed. The practical criteria for KDD projects are similar to those for other applications of advanced technology and include the potential impact of an application, the absence of simpler alternative solutions, and strong organizational support for using technology. For application dealing with personal data, one should also consider the privacy and legal issues (Piatetsky-Shapiro, 1995). The technical criteria include considerations such as the availability of sufficient data(cases).

This book seeks to provide the latest research and the best practices in the field of data mining. At the same-time it gives an in-depth look into the various emerging applications in data mining. Furthermore, this book provides an overview on the main issues of data mining (temporal association rule mining, classifiers, integration etc.), various new concepts in data mining and the application of data mining in various fields.

WHERE THE BOOK STANDS

In the global context, advanced mining techniques in data mining are important, especially in the realm of emerging domains. Data mining techniques are the outcome of an extensive process of study, research and product development. In essence, this evolution began when entrepreneurs started arching business data in computers, efforts continued with improvements in easier data access, and more recently researches generated technologies that allow users to navigate through their data in real time. Well developed information and communication network infrastructure and knowledge building applications in emerging domains are important to promote information exchange among users, data analysts, system
developers, and data mining researchers to facilitate the advances available from data mining research, application development, and technology transfer.

Advanced data mining techniques in emerging domains can yield substantial knowledge from even raw data that are primarily gathered for wider range of applications. The primary objective of the book is to develop various advanced data mining techniques and algorithms in emerging domains. Research in the field of data mining and knowledge discovery to evolve rapid and efficient ways of archiving and treating data has become a major field of study of mining in emerging domains.

Eventually, data mining is becoming a significant tool in science, engineering, industrial processes, healthcare, medicine, and other social services for making intelligent decision. However the datasets in these fields are predominantly large, complex, and often noisy. Therefore, extracting knowledge from data requires the use of sophisticated, high-performance and principled analysis techniques and algorithms, based on sound statistical foundations. These techniques in turn necessitate powerful visualization technologies; implementations that must be carefully tuned for performance; software systems that are usable by scientists, engineers, and physicians as well as researchers; and infrastructure that support them.

In this context, this book provides an overview on the main issues of data mining (temporal association rule mining, classifiers, integration etc.), various new concepts in data mining like disguised missing data, persistent strong rules, opinion mining, internet forums, and the application of data mining in the fields like telecommunication systems, biology, mobile marketing, microarrays and so forth.

**ORGANIZATION OF CHAPTERS**

This book includes fifteen chapters and they are divided into three sections: Concepts, Tools and Techniques; Research and Learning; and Case Studies. Section I has five chapters, and they illustrate the various tools and techniques for various emerging applications of data mining. Section II has five chapters, and they discuss policy and decision-making approaches of data mining for the development of business in terms of Intelligence and Marketing. The third section has five chapters and these chapters show various case studies on various trends and new domains of data mining applications.

In Chapter 1, the authors have explained the basic visual methods to detect some forms of disguises and the frameworks to identify them without requiring any domain expert. The chapter recognizes a data quality problem is that of disguised missing data which arises when an explicit code for missing data such as NA (Not Available) is not provided and a legitimate data value is used instead. Presence of these values may affect the outcome of data mining tasks severely such that association mining algorithms or clustering techniques may result in biased inaccurate association rules and invalid clusters respectively. Detection and elimination of these values are necessary but burdensome to be carried out manually. In this chapter, the methods to detect disguised missing values by visual inspection are explained first. Then, the authors describe the methods used to detect these values automatically. Finally, the framework to detect disguised missing data is proposed and a demonstration of the framework on spatial and categorical data sets is provided.

Chapter 2 provides a microarray technology tool analyze thousands of gene expression values with a single experiment. Authors state that due to the huge amount of data, most of recent studies are focused on the analysis and the extraction of useful and interesting information from microarray data. They also provide examples of applications which includes detecting genes highly correlated to diseases, selecting genes which show a similar behavior under specific conditions, building models to predict the disease
outcome based on genetic profiles, and inferring regulatory networks. This chapter presents a review of four popular data mining techniques (i.e., Classification, Feature Selection, Clustering, and Association Rule Mining) applied to microarray data. It describes the main characteristics of microarray data in order to understand the critical issues which are introduced by gene expression values analysis. Each technique is analyzed and examples of pertinent literature are reported. Furthermore, it provides the prospects of data mining research on microarray data.

Chapter 3 proposes a technique that is developed to explore frequent temporal itemsets in the database. The basic idea of this technique is to first partition the database into sub-databases in light of either common starting time or common ending time. Then for each partition, the proposed technique is used progressively to accumulate the number of occurrences of each candidate 2-itemsets. A Directed graph is built using the support of these candidate 2-itemsets (combined from all the sub-databases) as a result of generating all candidate temporal k-itemsets in the database. Therefore, the technique used in this chapter may help researchers not only to understand about generating frequent large temporal itemsets but also helps in understanding of finding temporal association rules among transactions within relational databases.

Chapter 4 discusses and reports how one can be benefited by using Data Mining and Knowledge Discovery techniques in achieving an acceptable level of quality of service of telecommunication systems. The quality of service is defined as the metrics which is predicated by using the data mining techniques, decision tree, association rules and neural networks. It further states that digital telecommunication networks are highly complex systems and thus their planning, management, and optimization are challenging tasks. The user expectations constitute the Quality of Service (QoS). It also states that to gain a competitive edge on other operators, the operating personnel have to measure the network in terms of QoS.

Chapter 5 demonstrates a process to identify especially powerful rules. More specifically, this chapter focuses on using association rules and classification mining to select the persistently strong association rules. Persistently strong association rules are association rules that are verifiable by classification mining the same data set. Further, the process for finding persistent strong rules was executed against two data sets obtained from the American National Election Studies. Analysis of the first data set resulted in one persistent strong rule and one persistent rule, while analysis of the second data set resulted in 11 persistent strong rules and 10 persistent rules. This chapter further suggests these rules are the most robust, consistent, and noteworthy among the much larger potential rule sets.

Chapter 6 discusses the relevance of Data Mining (DM) integration with Business Intelligence (BI), and its importance to business users. From the literature review, it was observed that the definition of an underlying structure for BI is missing, and therefore a framework is presented in this chapter. It was also observed that some efforts are being done that seek the establishment of standards in the DM field, both by academics and by people in the industry. Supported by those findings, this chapter introduces an architecture that can conduct to an effective usage of DM in BI. It also includes a DM language that is iterative and interactive in nature. This chapter suggests that the effective usage of DM in BI can be achieved by making DM models accessible to business users, through the use of the presented DM language.

Chapter 7 proposes Shadow Sensitive SWIFT commit protocol for Distributed Real Time Database Systems (DRTDBS), where only abort dependent cohort having deadline beyond a specific value (Tshadow_creation_time) can fork off a replica of itself called a shadow, whenever it borrows dirty value of a data item. It defines the new dependencies Commit-on-Termination external dependency between final commit operations of lender and shadow of its borrower and Begin-on-Abort internal dependency.
between shadow of borrower and borrower itself. The performance of Shadow Sensitive SWIFT is compared with shadow PROMPT, SWIFT and DSS-SWIFT commit protocols for both main memory resident and disk resident databases with and without communication delay. The chapter also shows that the proposed protocol improves the system performance up to 5% as transaction miss percentage.

In recent times, competition among mobile phone operators is now focused on switching customers away from competitors with extremely discounted telephony rates. In particular, this fierce competitive environment is the result of a saturated market with small or inexistent growth and has caused operators to rely increasingly on Value-Added Services (VAS) for revenue growth. Though mobile phone operators have thousands of different services available to offer to their customers, the contact opportunities to offer these services are limited. In this context, statistical methods and data mining tools can play an important role to optimize content delivery. In Chapter 8 the authors describe novel methods now available to mobile phone operators to optimize targeting and improve profitability from VAS offers.

Chapter 9 looks at various System Execution Modeling (SEM) tools which enable distributed system testers to validate Quality-of-Service (QoS) properties, such as end-to-end response time, throughput, and scalability, during early phases of the software lifecycle. Analytical capabilities of QoS properties, however, are traditionally bounded by a SEM tool’s capabilities. This chapter discusses how to mine system execution traces, which are a collection of log messages describing events and states of a distributed system throughout its execution lifetime, generated by distributed systems so that the validation of QoS properties is not dependent on a SEM tool’s capabilities. It also uses a real-life case study to illustrate how data mining system execution traces can assist in discovering potential performance bottlenecks using system execution traces.

Chapter 10 highlights how each one profits from the other and illustrates their cooperation in existing systems developed in the medical domain. Through a study, the authors have identified different types of cooperation that combine elicitation and data mining for knowledge acquisition, use expert knowledge to enact the knowledge discovery, use discovered knowledge to validate expert knowledge, and use discovered knowledge to improve the usability of an expert system. The chapter also describes the authors experience in combining expert and discovered knowledge in the development of a system for processing medical isokinetics data.

Chapter 11 focuses on discovering how Mesenchymal Stem Cells (MSCs) can be differentiated is an important topic in stem cell therapy and tissue engineering. In a general context, such differentiation analysis can be modeled as a classification problem in data mining. Specifically, this is concerned with the single-label multi-class classification task. The main aim of this chapter is to compare the performance of different associative classifiers, in terms of classification accuracy, efficiency, number of rules to be generated, quality of such rules, and the maximum number of attributes in rule-antecedents, with respect to MSC differentiation analysis.

Chapter 12 considers knowledge as a strategic weapon to get success in any business. Span of modern business applications have increased from a specific geographical area to the global world. The necessary resources of the business are available in distributed fashion using platform / technology like world wide web and grid of computational facilities. The prime intention of the grid architecture is to utilize scarce resources in objective to efficiently mine information from distributed resources. This chapter describes and differentiates World Wide Web (WWW), Semantic Web, Data Grid, and Knowledge Grid with the literature survey. Considering the limitations of the existing approaches, a generic multilayer architecture is designed and described with detailed methodology for each layer. The chapter also presents fuzzy XML technique to represent domain and meta knowledge into the knowledge repositories. To
In Chapter 13, the authors have introduced the concept “Opinion Mining” which is an emerging field of research concerned with applying computational methods to the treatment of subjectivity in text, with a number of applications in fields such as recommendation systems, contextual advertising and business intelligence. In this chapter the authors survey the area of opinion mining and discuss the SentiWordNet lexicon of sentiment information for terms derived from WordNet. Furthermore, the results of their research in applying this lexicon to sentiment classification of film reviews along with a novel approach that leverages opinion lexicons to build a data set of features used as input to a supervised learning classifier are also presented.

It is observed that the volume of information derived from post genomic technologies is rapidly increasing. Due to the amount of involved data, novel computational methods are needed for the analysis and knowledge discovery into the massive data sets produced by these new technologies. Furthermore, data integration is also gaining attention for merging signals from different sources in order to discover unknown relations. Chapter 14 presents a pipeline for biological data integration and discovery of a priori unknown relationships between gene expressions and metabolite accumulations. In this pipeline, two standard clustering methods are compared against a novel neural network approach. The neural model provides a simple visualization interface for identification of coordinated patterns variations, independently of the number of produced clusters. Moreover, the chapter proposes a method for the evaluation of the biological significance of the clusters found.

Finally, in Chapter 15, the authors focus on Internet forum, a web application for publishing user-generated content under the form of a discussion. Messages posted to the Internet forum form threads of discussion and contain textual and multimedia contents. Moreover the chapter addresses an important feature of Internet forums is their social aspect. Internet forums attract dedicated users who build tight social communities. The chapter discusses the architecture of Internet forums, presents an overview of data volumes involved and outlines technical challenges of scraping Internet forum data. A broad summary of all research conducted on mining and exploring Internet forums for social role discovery is also presented in this chapter.

CONCLUSION

Recent progress in scientific and engineering applications has accumulated huge volumes of high-dimensional data, stream data, and spatial and temporal data. Highly scalable and sophisticated data mining tools for such applications represent are of the most active research frontiers in data mining. Emerging applications like stream data, moving object data, RFID data, data from sensor networks, multi-agent data, semantic web, web search, biomedical engineering, telecommunications, geospatial data, climate data and Earth’s ecosystems etc. involve great management challenges that also represent new opportunities for data mining research.

In light of the tremendous amount of fast-growing and sophisticated types of data and comprehensive data analysis tasks, data mining technology may be only in its infancy, as the technology is still far from adequate for handling the large-scale and complex emerging application problems. Research is needed to develop highly automated, scalable, integrated reliable data mining systems and tools. Moreover, it is important to promote information exchange among users, data analysts, system developers, and data
mining researchers to facilitate the advances available from data mining research, application development and technology transfer.

The book incorporates various advanced data mining techniques and algorithms in emerging domains. Implementation of advanced data mining techniques will be very helpful for scientists, engineers, and physicians as well as researchers, and infrastructures that support them. In addition to these, as researchers revealed data mining is the process of automatic discovery of patterns, transformations, associations and anomalies in massive databases, and is a highly interdisciplinary field representing the confluence of multiple disciplines, such as database systems, data warehousing, machine learning, statistics, algorithms, data visualization, and high-performance computing. Utilizing advanced data mining techniques in emerging domains like stream data mining, mining multi-agent data, mining semantic web, ubiquitous knowledge discovery etc. can improve data mining process.

REFERENCES


