Introduction to Data Mining

SECOND EDITION

Pang-Ning Tan • Michael Steinbach • Anuj Karpatne • Vipin Kumar

INTRODUCTION TO DATA MINING

INTRODUCTION TO DATA MINING

SECOND EDITION

GLOBAL EDITION

PANG-NING TAN
Michigan State University

MICHAEL STEINBACH
University of Minnesota

ANUJ KARPATNE
University of Minnesota

VIPIN KUMAR
University of Minnesota

330 Hudson Street, NY NY 10013

Director, Portfolio Management: Engineering, Computer Science & Global Editions: Julian Partridge
Specialist, Higher Ed Portfolio Management: Matt Goldstein
Portfolio Management Assistant: Meghan Jacoby
Acquisitions Editor, Global Edition: Sourabh Maheshwari
Managing Content Producer: Scott Disanno
Content Producer: Carole Snyder
Senior Project Editor, Global Edition: K.K. Neelakantan
Web Developer: Steve Wright
Manager, Media Production, Global Edition: Vikram Kumar
Rights and Permissions Manager: Ben Ferrini
Manufacturing Buyer, Higher Ed, Lake Side Communications Inc (LSC): Maura Zaldivar-Garcia
Senior Manufacturing Controller, Global Edition: Caterina Pellegrino
Inventory Manager: Ann Lam
Product Marketing Manager: Yvonne Vannatta
Field Marketing Manager: Demetrius Hall
Marketing Assistant: Jon Bryant
Cover Designer: Lumina Datamatics
Full-Service Project Management: Ramya Radhakrishnan, Integra Software Services

Pearson Education Limited
KAO Two
KAO Park
Harlow
CM17 9NA
United Kingdom

and Associated Companies throughout the world

Visit us on the World Wide Web at: www.pearsonglobaleditions.com

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN 10: 0-273-76922-7
ISBN 13: 978-0-273-76922-4
eBook ISBN 13: 978-0-273-77532-4

eBook formatted by Integra Software Services.

To our families ...

Preface to the Second Edition

Since the first edition, roughly 12 years ago, much has changed in the field of data analysis. The volume and variety of data being collected continues to increase, as has the rate (velocity) at which it is being collected and used to make decisions. Indeed, the term Big Data has been used to refer to the massive and diverse data sets now available. In addition, the term data science has been coined to describe an emerging area that applies tools and techniques from various fields, such as data mining, machine learning, statistics, and many others, to extract actionable insights from data, often big data.

The growth in data has created numerous opportunities for all areas of data analysis. The most dramatic developments have been in the area of predictive modeling, across a wide range of application domains. For instance, recent advances in neural networks, known as deep learning, have shown impressive results in a number of challenging areas, such as image classification, speech recognition, as well as text categorization and understanding. While not as dramatic, other areas, e.g., clustering, association analysis, and anomaly detection have also continued to advance. This new edition is in response to those advances.

Overview As with the first edition, the second edition of the book provides a comprehensive introduction to data mining and is designed to be accessible and useful to students, instructors, researchers, and professionals. Areas covered include data preprocessing, predictive modeling, association analysis, cluster analysis, anomaly detection, and avoiding false discoveries. The goal is to present fundamental concepts and algorithms for each topic, thus providing the reader with the necessary background for the application of data mining to real problems. As before, classification, association analysis and cluster analysis, are each covered in a pair of chapters. The introductory chapter covers basic concepts, representative algorithms, and evaluation techniques, while the more following chapter discusses advanced concepts and algorithms. As before, our objective is to provide the reader with a sound understanding of the foundations of data mining, while still covering many important advanced

topics. Because of this approach, the book is useful both as a learning tool and as a reference.

To help readers better understand the concepts that have been presented, we provide an extensive set of examples, figures, and exercises. The solutions to the original exercises, which are already circulating on the web, will be made public. The exercises are mostly unchanged from the last edition, with the exception of new exercises in the chapter on avoiding false discoveries. New exercises for the other chapters and their solutions will be available to instructors via the web. Bibliographic notes are included at the end of each chapter for readers who are interested in more advanced topics, historically important papers, and recent trends. These have also been significantly updated. The book also contains a comprehensive subject and author index.

What is New in the Second Edition? Some of the most significant improvements in the text have been in the two chapters on classification. The introductory chapter uses the decision tree classifier for illustration, but the discussion on many topics—those that apply across all classification approaches—has been greatly expanded and clarified, including topics such as overfitting, underfitting, the impact of training size, model complexity, model selection, and common pitfalls in model evaluation. Almost every section of the advanced classification chapter has been significantly updated. The material on Bayesian networks, support vector machines, and artificial neural networks has been significantly expanded. We have added a separate section on deep networks to address the current developments in this area. The discussion of evaluation, which occurs in the section on imbalanced classes, has also been updated and improved.

The changes in association analysis are more localized. We have completely reworked the section on the evaluation of association patterns (introductory chapter), as well as the sections on sequence and graph mining (advanced chapter). Changes to cluster analysis are also localized. The introductory chapter added the K-means initialization technique and an updated the discussion of cluster evaluation. The advanced clustering chapter adds a new section on spectral graph clustering. Anomaly detection has been greatly revised and expanded. Existing approaches—statistical, nearest neighbor/density-based, and clustering based—have been retained and updated, while new approaches have been added: reconstruction-based, one-class classification, and information-theoretic. The reconstruction-based approach is illustrated using autoencoder networks that are part of the deep learning paradigm. The data chapter has

been updated to include discussions of mutual information and kernel-based techniques.

The last chapter, which discusses how to avoid false discoveries and produce valid results, is completely new, and is novel among other contemporary textbooks on data mining. It supplements the discussions in the other chapters with a discussion of the statistical concepts (statistical significance, p-values, false discovery rate, permutation testing, etc.) relevant to avoiding spurious results, and then illustrates these concepts in the context of data mining techniques. This chapter addresses the increasing concern over the validity and reproducibility of results obtained from data analysis. The addition of this last chapter is a recognition of the importance of this topic and an acknowledgment that a deeper understanding of this area is needed for those analyzing data.

The data exploration chapter has been deleted, as have the appendices, from the print edition of the book, but will remain available on the web. A new appendix provides a brief discussion of scalability in the context of big data.

To the Instructor As a textbook, this book is suitable for a wide range of students at the advanced undergraduate or graduate level. Since students come to this subject with diverse backgrounds that may not include extensive knowledge of statistics or databases, our book requires minimal prerequisites. No database knowledge is needed, and we assume only a modest background in statistics or mathematics, although such a background will make for easier going in some sections. As before, the book, and more specifically, the chapters covering major data mining topics, are designed to be as self-contained as possible. Thus, the order in which topics can be covered is quite flexible. The core material is covered in chapters 2 (data), 3 (classification), 4 (association analysis), 5 (clustering), and 9 (anomaly detection). We recommend at least a cursory coverage of Chapter 10 (Avoiding False Discoveries) to instill in students some caution when interpreting the results of their data analysis. Although the introductory data chapter (2) should be covered first, the basic classification (3), association analysis (4), and clustering chapters (5), can be covered in any order. Because of the relationship of anomaly detection (9) to classification (3) and clustering (5), these chapters should precede Chapter 9. Various topics can be selected from the advanced classification, association analysis, and clustering chapters (6, 7, and 8, respectively) to fit the schedule and interests of the instructor and students. We also advise that the lectures be augmented by projects or practical exercises in data mining. Although they

are time consuming, such hands-on assignments greatly enhance the value of the course.

Support Materials Support materials available to all readers of this book are available on the book's website.

- PowerPoint lecture slides
- Suggestions for student projects
- Data mining resources, such as algorithms and data sets
- Online tutorials that give step-by-step examples for selected data mining techniques described in the book using actual data sets and data analysis software

Additional support materials, including solutions to exercises, are available only to instructors adopting this textbook for classroom use.

Acknowledgments Many people contributed to the first and second editions of the book. We begin by acknowledging our families to whom this book is dedicated. Without their patience and support, this project would have been impossible.

We would like to thank the current and former students of our data mining groups at the University of Minnesota and Michigan State for their contributions. Eui-Hong (Sam) Han and Mahesh Joshi helped with the initial data mining classes. Some of the exercises and presentation slides that they created can be found in the book and its accompanying slides. Students in our data mining groups who provided comments on drafts of the book or who contributed in other ways include Shyam Boriah, Haibin Cheng, Varun Chandola, Eric Eilertson, Levent Ertöz, Jing Gao, Rohit Gupta, Sridhar Iyer, Jung-Eun Lee, Benjamin Mayer, Aysel Ozgur, Uygar Oztekin, Gaurav Pandey, Kashif Riaz, Jerry Scripps, Gyorgy Simon, Hui Xiong, Jieping Ye, and Pusheng Zhang. We would also like to thank the students of our data mining classes at the University of Minnesota and Michigan State University who worked with early drafts of the book and provided invaluable feedback. We specifically note the helpful suggestions of Bernardo Craemer, Arifin Ruslim, Jamshid Vayghan, and Yu Wei.

Joydeep Ghosh (University of Texas) and Sanjay Ranka (University of Florida) class tested early versions of the book. We also received many useful suggestions directly from the following UT students: Pankaj Adhikari, Rajiv Bhatia, Frederic Bosche, Arindam Chakraborty, Meghana Deodhar, Chris Everson, David Gardner, Saad Godil, Todd Hay, Clint Jones, Ajay Joshi,

Joonsoo Lee, Yue Luo, Anuj Nanavati, Tyler Olsen, Sunyoung Park, Aashish Phansalkar, Geoff Prewett, Michael Ryoo, Daryl Shannon, and Mei Yang.

Ronald Kostoff (ONR) read an early version of the clustering chapter and offered numerous suggestions. George Karypis provided invaluable LaTeX assistance in creating an author index. Irene Moulitsas also provided assistance with LaTeX and reviewed some of the appendices. Musetta Steinbach was very helpful in finding errors in the figures.

We would like to acknowledge our colleagues at the University of Minnesota and Michigan State who have helped create a positive environment for data mining research. They include Arindam Banerjee, Dan Boley, Joyce Chai, Anil Jain, Ravi Janardan, Rong Jin, George Karypis, Claudia Neuhauser, Haesun Park, William F. Punch, György Simon, Shashi Shekhar, and Jaideep Srivastava. The collaborators on our many data mining projects, who also have our gratitude, include Ramesh Agrawal, Maneesh Bhargava, Steve Cannon, Alok Choudhary, Imme Ebert-Uphoff, Auroop Ganguly, Piet C. de Groen, Fran Hill, Yongdae Kim, Steve Klooster, Kerry Long, Nihar Mahapatra, Rama Nemani, Nikunj Oza, Chris Potter, Lisiane Pruinelli, Nagiza Samatova, Jonathan Shapiro, Kevin Silverstein, Brian Van Ness, Bonnie Westra, Nevin Young, and Zhi-Li Zhang.

The departments of Computer Science and Engineering at the University of Minnesota and Michigan State University provided computing resources and a supportive environment for this project. ARDA, ARL, ARO, DOE, NASA, NOAA, and NSF provided research support for Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. In particular, Kamal Abdali, Mitra Basu, Dick Brackney, Jagdish Chandra, Joe Coughlan, Michael Coyle, Stephen Davis, Frederica Darema, Richard Hirsch, Chandrika Kamath, Tsengdar Lee, Raju Namburu, N. Radhakrishnan, James Sidoran, Sylvia Spengler, Bhavani Thuraisingham, Walt Tiernin, Maria Zemankova, Aidong Zhang, and Xiaodong Zhang have been supportive of our research in data mining and high-performance computing.

It was a pleasure working with the helpful staff at Pearson Education. In particular, we would like to thank Matt Goldstein, Kathy Smith, Carole Snyder, and Joyce Wells. We would also like to thank George Nichols, who helped with the art work and Paul Anagnostopoulos, who provided LaTeX support.

We are grateful to the following Pearson reviewers: Leman Akoglu (Carnegie Mellon University), Chien-Chung Chan (University of Akron), Zhengxin Chen (University of Nebraska at Omaha), Chris Clifton (Purdue University), Joydeep Ghosh (University of Texas, Austin), Nazli Goharian (Illinois Institute of Technology), J. Michael Hardin (University of Alabama), Jingrui He (Arizona

State University), James Hearne (Western Washington University), Hillol Kargupta (University of Maryland, Baltimore County and Agnik, LLC), Eamonn Keogh (University of California-Riverside), Bing Liu (University of Illinois at Chicago), Mariofanna Milanova (University of Arkansas at Little Rock), Srinivasan Parthasarathy (Ohio State University), Zbigniew W. Ras (University of North Carolina at Charlotte), Xintao Wu (University of North Carolina at Charlotte), and Mohammed J. Zaki (Rensselaer Polytechnic Institute).

Over the years since the first edition, we have also received numerous comments from readers and students who have pointed out typos and various other issues. We are unable to mention these individuals by name, but their input is much appreciated and has been taken into account for the second edition.

Acknowledgments for the Global Edition Pearson would like to thank and acknowledge Pramod Kumar Singh (Atal Bihari Vajpayee Indian Institute of Information Technology and Management) for contributing to the Global Edition, and Annappa (National Institute of Technology Surathkal), Komal Arora, and Soumen Mukherjee (RCC Institute of Technology) for reviewing the Global Edition.

Contents

This page is intentionally left blank

1

Introduction

Rapid advances in data collection and storage technology, coupled with the ease with which data can be generated and disseminated, have triggered the explosive growth of data, leading to the current age of **big data**. Deriving actionable insights from these large data sets is increasingly important in decision making across almost all areas of society, including business and industry; science and engineering; medicine and biotechnology; and government and individuals. However, the amount of data (volume), its complexity (variety), and the rate at which it is being collected and processed (velocity) have simply become too great for humans to analyze unaided. Thus, there is a great need for automated tools for extracting useful information from the big data despite the challenges posed by its enormity and diversity.

Data mining blends traditional data analysis methods with sophisticated algorithms for processing this abundance of data. In this introductory chapter, we present an overview of data mining and outline the key topics to be covered in this book. We start with a description of some applications that require more advanced techniques for data analysis.

Business and Industry Point-of-sale data collection (bar code scanners, radio frequency identification (RFID), and smart card technology) have allowed retailers to collect up-to-the-minute data about customer purchases at the checkout counters of their stores. Retailers can utilize this information, along with other business-critical data, such as web server logs from e-commerce websites and customer service records from call centers, to help them better understand the needs of their customers and make more informed business decisions.

Data mining techniques can be used to support a wide range of business intelligence applications, such as customer profiling, targeted marketing,

workflow management, store layout, fraud detection, and automated buying and selling. An example of the last application is high-speed stock trading, where decisions on buying and selling have to be made in less than a second using data about financial transactions. Data mining can also help retailers answer important business questions, such as "Who are the most profitable customers?"; "What products can be cross-sold or up-sold?"; and "What is the revenue outlook of the company for next year?" These questions have inspired the development of such data mining techniques as association analysis (Chapters 4 and 7).

As the Internet continues to revolutionize the way we interact and make decisions in our everyday lives, we are generating massive amounts of data about our online experiences, e.g., web browsing, messaging, and posting on social networking websites. This has opened several opportunities for business applications that use web data. For example, in the e-commerce sector, data about our online viewing or shopping preferences can be used to provide personalized recommendations of products. Data mining also plays a prominent role in supporting several other Internet-based services, such as filtering spam messages, answering search queries, and suggesting social updates and connections. The large corpus of text, images, and videos available on the Internet has enabled a number of advancements in data mining methods, including deep learning, which is discussed in Chapter 6. These developments have led to great advances in a number of applications, such as object recognition, natural language translation, and autonomous driving.

Another domain that has undergone a rapid big data transformation is the use of mobile sensors and devices, such as smart phones and wearable computing devices. With better sensor technologies, it has become possible to collect a variety of information about our physical world using low-cost sensors embedded on everyday objects that are connected to each other, termed the Internet of Things (IOT). This deep integration of physical sensors in digital systems is beginning to generate large amounts of diverse and distributed data about our environment, which can be used for designing convenient, safe, and energy-efficient home systems, as well as for urban planning of smart cities.

Medicine, Science, and Engineering Researchers in medicine, science, and engineering are rapidly accumulating data that is key to significant new discoveries. For example, as an important step toward improving our understanding of the Earth's climate system, NASA has deployed a series of Earth-orbiting satellites that continuously generate global observations of the land

surface, oceans, and atmosphere. However, because of the size and spatio-temporal nature of the data, traditional methods are often not suitable for analyzing these data sets. Techniques developed in data mining can aid Earth scientists in answering questions such as the following: "What is the relationship between the frequency and intensity of ecosystem disturbances such as droughts and hurricanes to global warming?"; "How is land surface precipitation and temperature affected by ocean surface temperature?"; and "How well can we predict the beginning and end of the growing season for a region?"

As another example, researchers in molecular biology hope to use the large amounts of genomic data to better understand the structure and function of genes. In the past, traditional methods in molecular biology allowed scientists to study only a few genes at a time in a given experiment. Recent breakthroughs in microarray technology have enabled scientists to compare the behavior of thousands of genes under various situations. Such comparisons can help determine the function of each gene, and perhaps isolate the genes responsible for certain diseases. However, the noisy, high-dimensional nature of data requires new data analysis methods. In addition to analyzing gene expression data, data mining can also be used to address other important biological challenges such as protein structure prediction, multiple sequence alignment, the modeling of biochemical pathways, and phylogenetics.

Another example is the use of data mining techniques to analyze electronic health record (EHR) data, which has become increasingly available. Not very long ago, studies of patients required manually examining the physical records of individual patients and extracting very specific pieces of information pertinent to the particular question being investigated. EHRs allow for a faster and broader exploration of such data. However, there are significant challenges since the observations on any one patient typically occur during their visits to a doctor or hospital and only a small number of details about the health of the patient are measured during any particular visit.

Currently, EHR analysis focuses on simple types of data, e.g., a patient's blood pressure or the diagnosis code of a disease. However, large amounts of more complex types of medical data are also being collected, such as electrocardiograms (ECGs) and neuroimages from magnetic resonance imaging (MRI) or functional Magnetic Resonance Imaging (fMRI). Although challenging to analyze, this data also provides vital information about patients. Integrating and analyzing such data, with traditional EHR and genomic data is one of the capabilities needed to enable precision medicine, which aims to provide more personalized patient care.

1.1 What Is Data Mining?

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large data sets in order to find novel and useful patterns that might otherwise remain unknown. They also provide the capability to predict the outcome of a future observation, such as the amount a customer will spend at an online or a brick-and-mortar store.

Not all information discovery tasks are considered to be data mining. Examples include queries, e.g., looking up individual records in a database or finding web pages that contain a particular set of keywords. This is because such tasks can be accomplished through simple interactions with a database management system or an information retrieval system. These systems rely on traditional computer science techniques, which include sophisticated indexing structures and query processing algorithms, for efficiently organizing and retrieving information from large data repositories. Nonetheless, data mining techniques have been used to enhance the performance of such systems by improving the quality of the search results based on their relevance to the input queries.

Data Mining and Knowledge Discovery in Databases

Data mining is an integral part of **knowledge discovery in databases (KDD)**, which is the overall process of converting raw data into useful information, as shown in Figure 1.1. This process consists of a series of steps, from data preprocessing to postprocessing of data mining results.

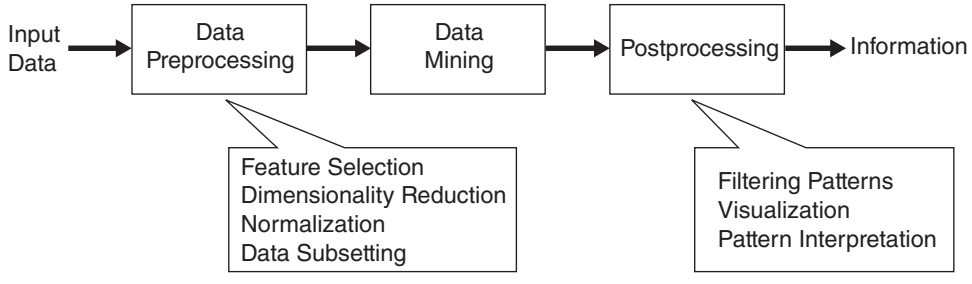

Figure 1.1. The process of knowledge discovery in databases (KDD).

The input data can be stored in a variety of formats (flat files, spreadsheets, or relational tables) and may reside in a centralized data repository or be distributed across multiple sites. The purpose of **preprocessing** is to transform the raw input data into an appropriate format for subsequent analysis. The steps involved in data preprocessing include fusing data from multiple sources, cleaning data to remove noise and duplicate observations, and selecting records and features that are relevant to the data mining task at hand. Because of the many ways data can be collected and stored, data preprocessing is perhaps the most laborious and time-consuming step in the overall knowledge discovery process.

"Closing the loop" is a phrase often used to refer to the process of integrating data mining results into decision support systems. For example, in business applications, the insights offered by data mining results can be integrated with campaign management tools so that effective marketing promotions can be conducted and tested. Such integration requires a **postprocessing** step to ensure that only valid and useful results are incorporated into the decision support system. An example of postprocessing is visualization, which allows analysts to explore the data and the data mining results from a variety of viewpoints. Hypothesis testing methods can also be applied during postprocessing to eliminate spurious data mining results. (See Chapter 10.)

1.2 Motivating Challenges

As mentioned earlier, traditional data analysis techniques have often encountered practical difficulties in meeting the challenges posed by big data applications. The following are some of the specific challenges that motivated the development of data mining.

Scalability Because of advances in data generation and collection, data sets with sizes of terabytes, petabytes, or even exabytes are becoming common. If data mining algorithms are to handle these massive data sets, they must be scalable. Many data mining algorithms employ special search strategies to handle exponential search problems. Scalability may also require the implementation of novel data structures to access individual records in an efficient manner. For instance, out-of-core algorithms may be necessary when processing data sets that cannot fit into main memory. Scalability can also be improved by using sampling or developing parallel and distributed algorithms. A general overview of techniques for scaling up data mining algorithms is given in Appendix F.

High Dimensionality It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago. In bioinformatics, progress in microarray technology has produced gene expression data involving thousands of features. Data sets with temporal or spatial components also tend to have high dimensionality. For example, consider a data set that contains measurements of temperature at various locations. If the temperature measurements are taken repeatedly for an extended period, the number of dimensions (features) increases in proportion to the number of measurements taken. Traditional data analysis techniques that were developed for low-dimensional data often do not work well for such high-dimensional data due to issues such as the curse of dimensionality (to be discussed in Chapter 2). Also, for some data analysis algorithms, the computational complexity increases rapidly as the dimensionality (the number of features) increases.

Heterogeneous and Complex Data Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical. As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes. Recent years have also seen the emergence of more complex data objects. Examples of such non-traditional types of data include web and social media data containing text, hyperlinks, images, audio, and videos; DNA data with sequential and three-dimensional structure; and climate data that consists of measurements (temperature, pressure, etc.) at various times and locations on the Earth's surface. Techniques developed for mining such complex objects should take into consideration relationships in the data, such as temporal and spatial autocorrelation, graph connectivity, and parent-child relationships between the elements in semi-structured text and XML documents.

Data Ownership and Distribution Sometimes, the data needed for an analysis is not stored in one location or owned by one organization. Instead, the data is geographically distributed among resources belonging to multiple entities. This requires the development of distributed data mining techniques. The key challenges faced by distributed data mining algorithms include the following: (1) how to reduce the amount of communication needed to perform the distributed computation, (2) how to effectively consolidate the data mining results obtained from multiple sources, and (3) how to address data security and privacy issues.

Non-traditional Analysis The traditional statistical approach is based on a hypothesize-and-test paradigm. In other words, a hypothesis is proposed, an experiment is designed to gather the data, and then the data is analyzed with respect to the hypothesis. Unfortunately, this process is extremely labor-intensive. Current data analysis tasks often require the generation and evaluation of thousands of hypotheses, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation. Furthermore, the data sets analyzed in data mining are typically not the result of a carefully designed experiment and often represent opportunistic samples of the data, rather than random samples.

1.3 The Origins of Data Mining

While data mining has traditionally been viewed as an intermediate process within the KDD framework, as shown in Figure 1.1, it has emerged over the years as an academic field within computer science, focusing on all aspects of KDD, including data preprocessing, mining, and postprocessing. Its origin can be traced back to the late 1980s, following a series of workshops organized on the topic of knowledge discovery in databases. The workshops brought together researchers from different disciplines to discuss the challenges and opportunities in applying computational techniques to extract actionable knowledge from large databases. The workshops quickly grew into hugely popular conferences that were attended by researchers and practitioners from both the academia and industry. The success of these conferences, along with the interest shown by businesses and industry in recruiting new hires with a data mining background, have fueled the tremendous growth of this field.

The field was initially built upon the methodology and algorithms that researchers had previously used. In particular, data mining researchers draw upon ideas, such as (1) sampling, estimation, and hypothesis testing from statistics and (2) search algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning. Data mining has also been quick to adopt ideas from other areas, including optimization, evolutionary computing, information theory, signal processing, visualization, and information retrieval, and extending them to solve the challenges of mining big data.

A number of other areas also play key supporting roles. In particular, database systems are needed to provide support for efficient storage, indexing,

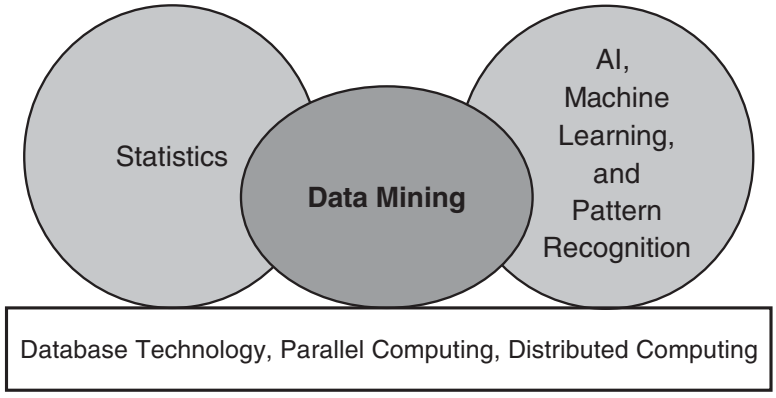

Figure 1.2. Data mining as a confluence of many disciplines.

and query processing. Techniques from high performance (parallel) computing are often important in addressing the massive size of some data sets. Distributed techniques can also help address the issue of size and are essential when the data cannot be gathered in one location. Figure 1.2 shows the relationship of data mining to other areas.

Data Science and Data-Driven Discovery

Data science is an interdisciplinary field that studies and applies tools and techniques for deriving useful insights from data. Although data science is regarded as an emerging field with a distinct identity of its own, the tools and techniques often come from many different areas of data analysis, such as data mining, statistics, AI, machine learning, pattern recognition, database technology, and distributed and parallel computing. (See Figure 1.2.)

The emergence of data science as a new field is a recognition that, often, none of the existing areas of data analysis provides a complete set of tools for the data analysis tasks that are often encountered in emerging applications. Instead, a broad range of computational, mathematical, and statistical skills is often required. To illustrate the challenges that arise in analyzing such data, consider the following example. Social media and the Web present new opportunities for social scientists to observe and quantitatively measure human behavior on a large scale. To conduct such a study, social scientists work with analysts who possess skills in areas such as web mining, natural language processing (NLP), network analysis, data mining, and statistics. Compared to more traditional research in social science, which is often based on surveys, this analysis requires a broader range of skills and tools, and involves far larger

amounts of data. Thus, data science is, by necessity, a highly interdisciplinary field that builds on the continuing work of many fields.

The data-driven approach of data science emphasizes the direct discovery of patterns and relationships from data, especially in large quantities of data, often without the need for extensive domain knowledge. A notable example of the success of this approach is represented by advances in neural networks, i.e., deep learning, which have been particularly successful in areas which have long proved challenging, e.g., recognizing objects in photos or videos and words in speech, as well as in other application areas. However, note that this is just one example of the success of data-driven approaches, and dramatic improvements have also occurred in many other areas of data analysis. Many of these developments are topics described later in this book.

Some cautions on potential limitations of a purely data-driven approach are given in the Bibliographic Notes.

1.4 Data Mining Tasks

Data mining tasks are generally divided into two major categories:

Predictive tasks The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the **target** or **dependent variable**, while the attributes used for making the prediction are known as the **explanatory** or **independent variables**.

Descriptive tasks Here, the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory in nature and frequently require postprocessing techniques to validate and explain the results.

Figure 1.3 illustrates four of the core data mining tasks that are described in the remainder of this book.

Predictive modeling refers to the task of building a model for the target variable as a function of the explanatory variables. There are two types of predictive modeling tasks: **classification**, which is used for discrete target variables, and **regression**, which is used for continuous target variables. For example, predicting whether a web user will make a purchase at an online bookstore is a classification task because the target variable is binary-valued.

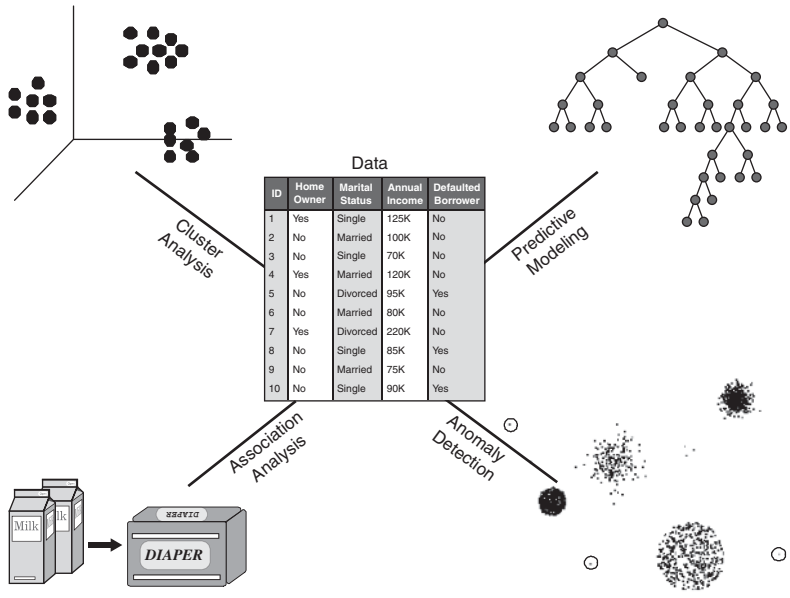

Figure 1.3. Four of the core data mining tasks.

On the other hand, forecasting the future price of a stock is a regression task because price is a continuous-valued attribute. The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable. Predictive modeling can be used to identify customers who will respond to a marketing campaign, predict disturbances in the Earth's ecosystem, or judge whether a patient has a particular disease based on the results of medical tests.

Example 1.1 (**Predicting the Type of a Flower**). Consider the task of predicting a species of flower based on the characteristics of the flower. In particular, consider classifying an Iris flower as one of the following three Iris species: Setosa, Versicolour, or Virginica. To perform this task, we need a data set containing the characteristics of various flowers of these three species. A data set with this type of information is the well-known Iris data set from the UCI Machine Learning Repository at `http://www.ics.uci.edu/~mlearn`. In addition to the species of a flower, this data set contains four other attributes: sepal width, sepal length, petal length, and petal width. Figure 1.4 shows a plot of petal width versus petal length for the 150 flowers in the Iris data set. Petal width is broken into the categories *low*, *medium*, and *high*, which correspond to the intervals [0, 0.75), [0.75, 1.75), [1.75, ∞), respectively. Also,