Preface

Advances in sensors, information technology, and high-performance computing have resulted in massive data sets becoming available in many scientific disciplines. These data sets are not only very large, being measured in terabytes and petabytes, but are also quite complex. This complexity arises as the data are collected by different sensors, at different times, at different frequencies, and at different resolutions. Further, the data are usually in the form of images or meshes, and often have both a spatial and a temporal component. These data sets arise in diverse fields such as astronomy, medical imaging, remote sensing, nondestructive testing, physics, materials science, and bioinformatics. They can be obtained from simulations, experiments, or observations.

This increasing size and complexity of data in scientific disciplines has resulted in a challenging problem. Many of the traditional techniques from visualization and statistics that were used for the analysis of these data are no longer suitable. Visualization techniques, even for moderate-sized data, are impractical due to their subjective nature and human limitations in absorbing detail, while statistical techniques do not scale up to massive data sets. As a result, much of the data collected are never even looked at, and the full potential of our advanced data collecting capabilities is only partially realized, if at all.

Data mining is the process concerned with uncovering patterns, associations, anomalies, and statistically significant structures in data. It is an iterative and interactive process involving data preprocessing, search for patterns, and visualization and validation of the results. It is a multidisciplinary field, borrowing and enhancing ideas from domains including image understanding, statistics, machine learning, mathematical optimization, high-performance computing, information retrieval, and computer vision. Data mining techniques hold the promise of assisting scientists and engineers in the analysis of massive, complex data sets, enabling them to make scientific discoveries, gain fundamental insights into the physical processes being studied, and advance their understanding of the world around us.

Why focus on scientific data?

There are several books available on data mining. Many focus on the specific task of finding patterns in the data through techniques such as decision trees or neural networks, while others focus on commercial data and are targeted at the business communities. Pattern recognition techniques are necessary in the mining of scientific data; however, they form only a part of the entire data mining process. Texts focusing on pattern recognition do not discuss how to convert raw scientific data in the form of images or meshes into a form that

can be used as input to a pattern recognition algorithm. On the other hand, texts on mining business applications focus on data that have been cleaned and are in a database, a situation that is rarely true for scientific data. Further, the assumptions made for business data may not hold for scientific data. For example, in a targeted marketing application, it is possible to use historical data to generate a large training set, with equal number of positive and negative examples. In contrast, training sets in scientific data tend to be quite small as they are generated manually. They are also frequently unbalanced, with far more examples of one type than another.

The data mining problems encountered in science and engineering applications may seem very different at first glance. However, there are several common threads among them. This book focuses on the identification of these common problems and their potential solutions. It considers the end-to-end process of data mining, starting with the raw data in the form of images or meshes, preprocessing the data, and finding useful information in the data that are then shown to the scientist for validation. There is greater focus on the pre- and postprocessing steps as these are often the more critical and time-consuming parts of mining scientific data. In the process, this book brings together topics that are often spread out among books focusing on image processing, information retrieval, or mathematical optimization. This book has also been written with an emphasis on the practical aspects of mining scientific data and introduces the reader to topics not often covered in academic texts such as data mining systems.

Any book must be limited in scope if it has any hope of seeing the light of day (or the printer's ink); therefore, several topics are beyond the scope of this book. These include:

- **Database technology:** In many scientific disciplines, metadata representing aspects of the data, such as the time it was collected and the circumstances under which it was collected, may be stored in a database. However, the data itself is usually stored in flat files, with perhaps a pointer from the associated metadata in the database. As a result, database technology is often not very relevant to scientific data. In fact, when many scientists use the term "database," they are referring to their data as a whole; these data may be in flat files if they are online or they may just be a collection of tapes in a file cabinet.

- **Online Analytical Processing (OLAP):** This includes tasks such as querying the database and extracting slices or cubes from the data that satisfy certain constraints. Such techniques are used in some scientific applications for tasks such as exploratory data analysis and subsetting the data for further analysis. Tools to support this are provided either through visualization software or by domain- or problem-specific software.

- **Parallel implementations:** While these are often necessary to make the problem of mining massive data sets tractable, the subject is broad enough that several books can be devoted to it.

- **Collection of the data:** The process used for collection of the data, whether from experiments, observations, or simulations, is often an important aspect of data mining. In particular, it can dictate the choice of algorithms to use as well as the confidence in the conclusions being drawn from the data. However, it is also a topic that is very dependent on both the problem and the application domain, and is therefore beyond the scope of this book.

- **Storage and access of the data:** While these are topics germane to mining scientific data, they are more appropriate for a text on data management.

In addition, since there are several excellent texts on pattern recognition algorithms, I will discuss this topic only briefly in Chapter 11, focusing on issues more relevant to practical applications and providing pointers as appropriate.

What is in this book?

This book is focused on the practical aspects of scientific data mining. Over the last ten years, as I analyzed data from various scientific domains, I encountered a diversity of problems. Often, the solutions were not found in standard texts on data mining, but in texts on different domains, or on techniques developed in the context of a different problem in a different application area. In many cases, I had to modify existing techniques or innovate and come up with new approaches which exploited domain-specific information. In the process, I realized that there are many pitfalls in mining scientific data, and one must apply techniques with care, while being aware of the characteristics of the data and the assumptions of the different algorithms. The end goal of finding scientifically meaningful information in the data cannot be achieved by simply applying techniques blindly to the data.

This book tries to bring together, in a coherent whole, the different techniques which can be used in solving a variety of problems in the analysis of data from various scientific domains. Not all the techniques are useful in all problems and some techniques developed in the context of one application domain may find use in an entirely different application domain.

The chapters in the book are divided into five broad areas.

- **Data mining in scientific applications:** The first two chapters focus on an introduction to data mining (Chapter 1) and the role it plays in science and engineering applications (Chapter 2).

- **The data mining process:** Chapter 3 identifies the common themes in the applications discussed in Chapter 2. It also describes the data types used in scientific applications—these are used to motivate the data mining process described in Chapter 4. This chapter discusses how we start from the raw data in the form of images or meshes and the steps we take to extract useful information from them.

- **The tasks in the data mining process:** Chapters 5 through 12 are devoted to discussing the specific tasks in the data mining process. Chapter 5 discusses various ways of reducing the size of the data through techniques such as sampling and multiresolution. Chapter 6 discusses data fusion, or how we can combine complementary data from different sources. Chapter 7 describes how to remove noise from data, followed by algorithms for identifying objects in the data in Chapter 8. Once the objects have been identified, there are various representative features that can be extracted from the objects. These are discussed in Chapter 9. At the end of this extraction step, we have a list of objects that have been identified in the data, with each object described by a number of features. Often, these features number in the tens or hundreds. Ways of reducing this number are discussed in Chapter 10. The original raw data have now been reduced to a form that can be input to various pattern recognition

algorithms, as described in Chapter 11. Finally, the patterns identified in the data are visualized for validation by the domain scientist, as discussed in Chapter 12. These same visualization techniques can also be used in validating the results from any task in the data mining process, as well as in the initial exploratory analysis of the data which is done to determine either the next task or an appropriate algorithm for a task.

In each of these chapters, the intent is to introduce the reader to some commonly used techniques for a particular task, as well as some of the more advanced techniques that have been proposed recently. While the pros and cons of the traditional approaches are well known, the recent techniques are still topics of active research. As appropriate, suggestions for further reading are recommended in each chapter.

The techniques which are listed for the different tasks are by no means exhaustive. These are just the techniques I have found useful based on my experiences thus far. The omission of a particular solution approach is indicative either of my ignorance, or the space and time limitations which accompany the writing of any book.

- **Data mining systems:** Chapter 13 describes several data mining systems that have been developed for science and engineering data. These illustrate the challenges faced in building an end-to-end system and the various approaches taken to address these challenges.

- **Lessons learned, challenges, and opportunities:** Chapter 14 summarizes, both from my personal perspective as well as those of other data miners, the lessons learned in mining scientific data. It also describes the challenges and opportunities that await a data miner who takes on the task of analyzing scientific data.

The chapters in this book have been written to be as self contained as possible, so that a reader can focus on a topic of interest without having to read the entire book from the beginning. However, it is my hope that the reader who perseveres through the entire book will come away with an appreciation of the technical challenges in scientific data mining, the opportunities to borrow ideas from other fields, and the potential for making exciting scientific discoveries.

In this book, I have included the URLs for any information available on the Web. While I realize that some of these links may be short-lived, I find it helpful to include the information, as a Web search can provide the new location if a link has changed. I did not want to take the alternative of excluding the links, as it meant that I could not refer to the wealth of information easily accessible to all via the Web.

Several of the diagrams appearing in this book have been inspired by similar diagrams appearing elsewhere, as follows:

- Figure 5.1 is derived from the article by Witkin [629].

- Figure 5.2 is derived from Figure 2.2 in the text by Lindeberg [382].

- Figure 5.3 is derived from Figure 2.1 in the text by Stollnitz, DeRose, and Salesin [568].

- Figure 7.2 is derived from the article by Smith and Brady [553].

- Figure 11.2 is derived from Figure 4.2 in the book by Mitchell [430].

Finally, despite my best efforts, I am sure there are some unintentional errors in the book, either typographical errors, ambiguous statements, or incorrect statements reflecting my less-than-thorough understanding of a topic. In the absence of a coauthor to whom I could attribute these errors, I assume full responsibility for them. The opinions presented in the book represent my experiences with scientific data mining thus far; these opinions may not necessarily reflect the views of my colleagues, funding sources, employers, or the domain scientists I have worked with over the years. As I continue my education in the many fields that comprise this exciting subject, it is very likely that my thoughts will evolve as well. I also acknowledge that the book is biased toward techniques and topics I have considered in addressing the problems I have worked on in the last decade. I am sure there are others I have overlooked in my ignorance. I will gratefully receive all errata as well as suggestions for improvement.

Acknowledgments

It is my great pleasure to acknowledge all those who contributed, directly or indirectly, to this book.

This book grew out of tutorials I presented at the SIAM International Conference on Data Mining in 2001 and 2003, as well as an overview tutorial at a week-long program I organized on Scientific Data Mining at the Institute for Pure and Applied Mathematics, University of California, Los Angeles, in 2002. I would like to thank Linda Thiel of SIAM for suggesting that I expand the course material presented in these tutorials. She helped me to get started with the logistics of writing the book, providing many ideas for improvement along the way. It has been my pleasure to work with her, Sara Murphy, and the many members of SIAM staff; I truly appreciate their support, encouragement, and patience over the years.

I was introduced to the subject of data mining by Shivakumar Vaithyanathan, who helped me to realize that not only was this topic very different from data bases, it also used some of the techniques I knew from numerical linear algebra. My education in the many fields which comprise scientific data mining continued through interactions with the members of my project, Sapphire, at the Lawrence Livermore National Laboratory. It was a wonderful experience to work with all of them, especially Erick Cantú-Paz, Samson Sen-Ching Cheung, and Charles R. Musick who contributed to many stimulating discussions. This book reflects many of the ideas and efforts of the Sapphire team members.

The Sapphire project was funded by several different sources from 1998 to 2008, including the ASC and SciDAC programs at the US Department of Energy, the LDRD program at Lawrence Livermore National Laboratory, and the US Department of Homeland Security. I am very grateful to these programs, and the individuals associated with them, for supporting my work and thus indirectly contributing to this book.

I gained much of my practical experience in scientific data mining by working with real data sets and solving real problems. I owe a significant set of debts to all the scientists who generously shared their data, their time, and their domain expertise with my project team. It was my good fortune (or perhaps serendipity) that the first data set we analyzed was the Faint Images of the Radio Sky at Twenty centimeters (FIRST) data. The FIRST astronomers—Robert H. Becker, Richard L. White, Michael D. Gregg, and Sally A. Laurent-Muehleisen—taught me much about scientific data analysis and provided a gentle introduction to collaborating with domain scientists. Many others provided a fascinating

glimpse into the "science" aspect of scientific data mining, including Charles Alcock, Joshua Breslau, Keith H. Burrell, William H. Cabot, Andrew W. Cook, John A. Futterman, Omar A. Hurricane, Scott A. Klasky, Zhihong Lin, Ricardo J. Maqueda, Alfredo A. Marchetti, Paul L. Miller, Donald A. Monticello, Neil Pomphrey, Benjamin D. Santer, Brian K. Spears, Daren P. Stotler, Frederick H. Streitz, Michael A. Walker, Nathan G. Wimer, and Stewart J. Zweben. These scientists gave me a wonderful opportunity to apply analysis techniques to a diverse set of challenging problems.

I owe a profound intellectual debt to those who contributed toward my education—my parents, my brother Jayant, my late brother Hemant, my husband Sisira, and my teachers at the Lady Irwin School in New Delhi, India; the Indian Institute of Technology, Bombay, India; and the University of Illinois at Urbana-Champaign, USA.

Finally, I would like to thank the three people who, perhaps unknowingly, contributed immensely to this book—my mother, Sharada, for sharing with me her love of mathematics; my late father, Taranath, for teaching me English grammar and writing skills (I hope I have not let him down!); and my husband, Sisira Weeratunga, for his incredible support over the years, his patience as this book took far longer to complete than anticipated, and the countless technical discussions, which were both stimulating and educational. I dedicate this book to them, in gratitude.

November 2008 *Chandrika Kamath*