Preface

The interest in writing this book began at the IEEE International Conference on Intelligence and Security Informatics held in Washington, DC (June 11–14, 2012), where Mr. Matthew Amboy, the editor of *Business and Economics: OR and MS*, published by Springer Science+Business Media, expressed the need for a book on this topic, mainly focusing on a topic in data science field. The interest went even deeper when I attended the workshop conducted by Professor Bin Yu (Department of Statistics, University of California, Berkeley) and Professor David Madigan (Department of Statistics, Columbia University) at the Institute for Mathematics and its Applications, University of Minnesota on June 16–29, 2013.

Data science is one of the emerging fields in the twenty-first century. This field has been created to address the big data problems encountered in the day-to-day operations of many industries, including financial sectors, academic institutions, information technology divisions, health care companies, and government organizations. One of the important big data problems that needs immediate attention is in big data classifications. The network intrusion detection, public space intruder detection, fraud detection, spam filtering, and forensic linguistics are some of the practical examples of big data classification problems that require immediate attention.

We need significant collaboration between the experts in many disciplines, including mathematics, statistics, computer science, engineering, biology, and chemistry to find solutions to this challenging problem. Educational resources, like books and software, are also needed to train students to be the next generation of research leaders in this emerging research field. One of the current fields that brings the interdisciplinary experts, educational resources, and modern technologies under one roof is machine learning, which is a subfield of artificial intelligence.

Many models and algorithms for standard classification problems are available in the machine learning literature. However, a few of them are suitable for big data classification. Big data classification is dependent not only on the mathematical and software techniques but also on the computer technologies that help store, retrieve, and process the data with efficient scalability, accessibility, and computability features. One such recent technology is the distributed file system. A particular system

that has become popular and provides these features is the Hadoop distributed file system, which uses the modern techniques called MapReduce programming model (or a framework) with Mapper and Reducer functions that adopt the concept called the (key, value) pairs. The machine learning techniques such as the decision tree (a hierarchical approach), random forest (an ensemble hierarchical approach), and deep learning (a layered approach) are highly suitable for the system that addresses big data classification problems. Therefore, the goal of this book is to present some of the machine learning models and algorithms, and discuss them with examples.

The general objective of this book is to help readers, especially students and newcomers to the field of big data and machine learning, to gain a quick understanding of the techniques and technologies; therefore, the theory, examples, and programs (Matlab and R) presented in this book have been simplified, hardcoded, repeated, or spaced for improvements. They provide vehicles to test and understand the complicated concepts of various topics in the field. It is expected that the readers adopt these programs to experiment with the examples, and then modify or write their own programs toward advancing their knowledge for solving more complex and challenging problems.

The presentation format of this book focuses on simplicity, readability, and dependability so that both undergraduate and graduate students as well as new researchers, developers, and practitioners in this field can easily trust and grasp the concepts, and learn them effectively. The goal of the writing style is to reduce the mathematical complexity and help the vast majority of readers to understand the topics and get interested in the field. This book consists of four parts, with a total of 14 chapters. Part I mainly focuses on the topics that are needed to help analyze and understand big data. Part II covers the topics that can explain the systems required for processing big data. Part III presents the topics required to understand and select machine learning techniques to classify big data. Finally, Part IV concentrates on the topics that explain the scaling-up machine learning, an important solution for modern big data problems.

Greensboro, NC, USA Shan Suthaharan

http://www.springer.com/978-1-4899-7640-6

Machine Learning Models and Algorithms for Big Data Classification
Thinking with Examples for Effective Learning
Suthaharan, S.
2016, XIX, 359 p. 149 illus., 82 illus. in color., Hardcover
ISBN: 978-1-4899-7640-6