Preface

The purpose of this book is to provide up-to-date progress both in Multiple Criteria Programming (MCP) and Support Vector Machines (SVMs) that have become powerful tools in the field of data mining. Most of the content in this book are directly from the research and application activities that our research group has conducted over the last ten years.

Although the data mining community is familiar with Vapnik's SVM [206] in classification, using optimization techniques to deal with data separation and data analysis goes back more than fifty years. In the 1960s, O.L. Mangasarian formulated the principle of large margin classifiers and tackled it using linear programming. He and his colleagues have reformed his approaches in SVMs [141]. In the 1970s, A. Charnes and W.W. Cooper initiated Data Envelopment Analysis, where linear or quadratic programming is used to evaluate the efficiency of decision-making units in a given training dataset. Started from the 1980s, F. Glover proposed a number of linear programming models to solve the discriminant problem with a small-size of dataset [75]. Since 1998, the author and co-authors of this book have not only proposed and extended such a series of optimization-based classification models via Multiple Criteria Programming (MCP), but also improved a number of SVM related classification methods. These methods are different from statistics, decision tree induction, and neural networks in terms of the techniques of separating data.

When MCP is used for classification, there are two common criteria. The first one is the overlapping degree (e.g., norms of all overlapping) with respect to the separating hyperplane. The lower this degree, the better the classification. The second is the distance from a point to the separating hyperplane. The larger the sum of these distances, the better the classification. Accordingly, in linear cases, the objective of classification is either minimizing the sum of all overlapping or maximizing the sum of the distances. MCP can also be viewed as extensions of SVM. Under the framework of mathematical programming, both MCP and SVM share the same advantage of using a hyperplane for separating the data. With certain interpretation, MCP measures all possible distances from the training samples to separating hyperplane, while SVM only considers a fixed distance from the support vectors. This allows MCP approaches to become an alternative for data separation.

As we all know, optimization lies at the heart of most data mining approaches. Whenever data mining problems, such as classification and regression, are formulated by MCP or SVM, they can be reduced into different types of optimization problems, including quadratic, linear, nonlinear, fuzzy, second-order cone, semi-definite, and semi-infinite programs.

This book mainly focuses on MCP and SVM, especially their recent theoretical progress and real-life applications in various fields. Generally speaking, the book is organized into three parts, and each part contains several related chapters. Part one addresses some basic concepts and important theoretical topics on SVMs. It contains Chaps. 1, 2, 3, 4, 5, and 6. Chapter 1 reviews standard C-SVM for classification problem and extends it to problems with nominal attributes. Chapter 2 introduces LOO bounds for several algorithms of SVMs, which can speed up the process of searching for appropriate parameters in SVMs. Chapters 3 and 4 consider SVMs for multi-class, unsupervised, and semi-supervised problems by different mathematical programming models. Chapter 5 describes robust optimization models for several uncertain problems. Chapter 6 combines standard SVMs with feature selection strategies at the same time via p-norm minimization where $0 < p < 1$.

Part two mainly deals with MCP for data mining. Chapter 7 first introduces basic concepts and models of MCP, and then constructs penalized Multiple Criteria Linear Programming (MCLP) and regularized MCLP. Chapters 8, 9 and 11 describe several extensions of MCLP and Multiple Criteria Quadratic Programming (MCQP) in order to build different models under various objectives and constraints. Chapter 10 provides non-additive measured MCLP when interactions among attributes are allowed for classification.

Part three presents a variety of real-life applications of MCP and SVMs models. Chapters 12, 13, and 14 are finance applications, including firm financial analysis, personal credit management and health insurance fraud detection. Chapters 15 and 16 are about web services, including network intrusion detection and the analysis for the pattern of lost VIP email customer accounts. Chapter 17 is related to HIV-1 informatics for designing specific therapies, while Chap. 18 handles antigen and anti-body informatics. Chapter 19 concerns geochemical analyses. For the convenience of the reader, each chapter of applications is self-contained and self-explained.

Finally, Chap. 20 introduces the concept of intelligent knowledge management first time and describes in detail the theoretical framework of intelligent knowledge. The contents of this chapter go beyond the traditional domain of data mining and look for how to produce knowledge support to the end users by combing hidden patterns from data mining and human knowledge.

We are indebted to many people around the work for their encouragement and kind support of our research on MCP and SVMs. We would like to thank Prof. Naiyang Deng (China Agricultural University), Prof. Wei-xuan Xu (Institute of Policy and Management, Chinese Academy of Sciences), Prof. Zhengxin Chen (University of Nebraska at Omaha), Prof. Ling-ling Zhang (Graduate University of Chinese Academy of Sciences), Dr. Chun-hua Zhang (RenMin University of China), Dr. Zhi-xia Yang (XinJiang University, China), and Dr. Kun Zhao (Beijing WuZi University).

In the last five years, there are a number of colleagues and graduate students at the Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences who contributed to our research projects as well as the preparation of this book. Among them, we want to thank Dr. Xiao-fei Zhou, Dr. Ling-feng Niu, Dr. Xing-sen Li, Dr. Peng Zhang, Dr. Dong-ling Zhang, Dr. Zhi-wang Zhang, Dr. Yue-jin Zhang, Zhan Zhang, Guang-li Nie, Ruo-ying Chen, Zhong-bin OuYang, Wen-jing Chen, Ying Wang, Yue-hua Zhang, Xiu-xiang Zhao, Rui Wang.

Finally, we would like acknowledge a number of funding agencies who provided their generous support to our research activities on this book. They are First Data Corporation, Omaha, USA for the research fund "Multiple Criteria Decision Making in Credit Card Portfolio Management" (1998); the National Natural Science Foundation of China for the overseas excellent youth fund "Data Mining in Bank Loan Risk Management" (#70028101, 2001–2003), the regular project "Multiple Criteria Non-linear Based Data Mining Methods and Applications" (#70472074, 2005–2007), the regular project "Convex Programming Theory and Methods in Data Mining" (#10601064, 2007–2009), the key project "Optimization and Data Mining" (#70531040, 2006–2009), the regular project "Knowledge-Driven Multi-criteria Decision Making for Data Mining: Theories and Applications" (#70901011, 2010–2012), the regular project "Towards Reliable Software: A Standardize for Software Defects Measurement & Evaluation" (#70901015, 2010–2012), the innovative group grant "Data Mining and Intelligent Knowledge Management" (#70621001, #70921061, 2007–2012); the President Fund of Graduate University of Chinese Academy of Sciences; the Global Economic Monitoring and Policy Simulation Pre-research Project, Chinese Academy of Sciences (#KACX1-YW-0906, 2009–2011); US Air Force Research Laboratory for the contract "Proactive and Predictive Information Assurance for Next Generation Systems (P2INGS)" (#F30602-03-C-0247, 2003–2005); Nebraska EPScOR, the National Science Foundation of USA for industrial partnership fund "Creating Knowledge for Business Intelligence" (2009–2010); BHP Billiton Co., Australia for the research fund "Data Mining for Petroleum Exploration" (2005–2010); Nebraska Furniture Market—a unit of Berkshire Hathaway Investment Co., Omaha, USA for the research fund "Revolving Charge Accounts Receivable Retrospective Analysis" (2008–2009); and the CAS/SAFEA International Partnership Program for Creative Research Teams "Data Science-Based Fictitious Economy and Environmental Policy Research" (2010–2012).

Chengdu, China
December 31, 2010

Yong Shi
Yingjie Tian
Gang Kou
Yi Peng
Jianping Li

http://www.springer.com/978-0-85729-503-3

Optimization Based Data Mining: Theory and Applications

Shi, Y.; Tian, Y.; Kou, G.; Peng, Y.; Li, J.

2011, XVI, 316 p., Hardcover
ISBN: 978-0-85729-503-3