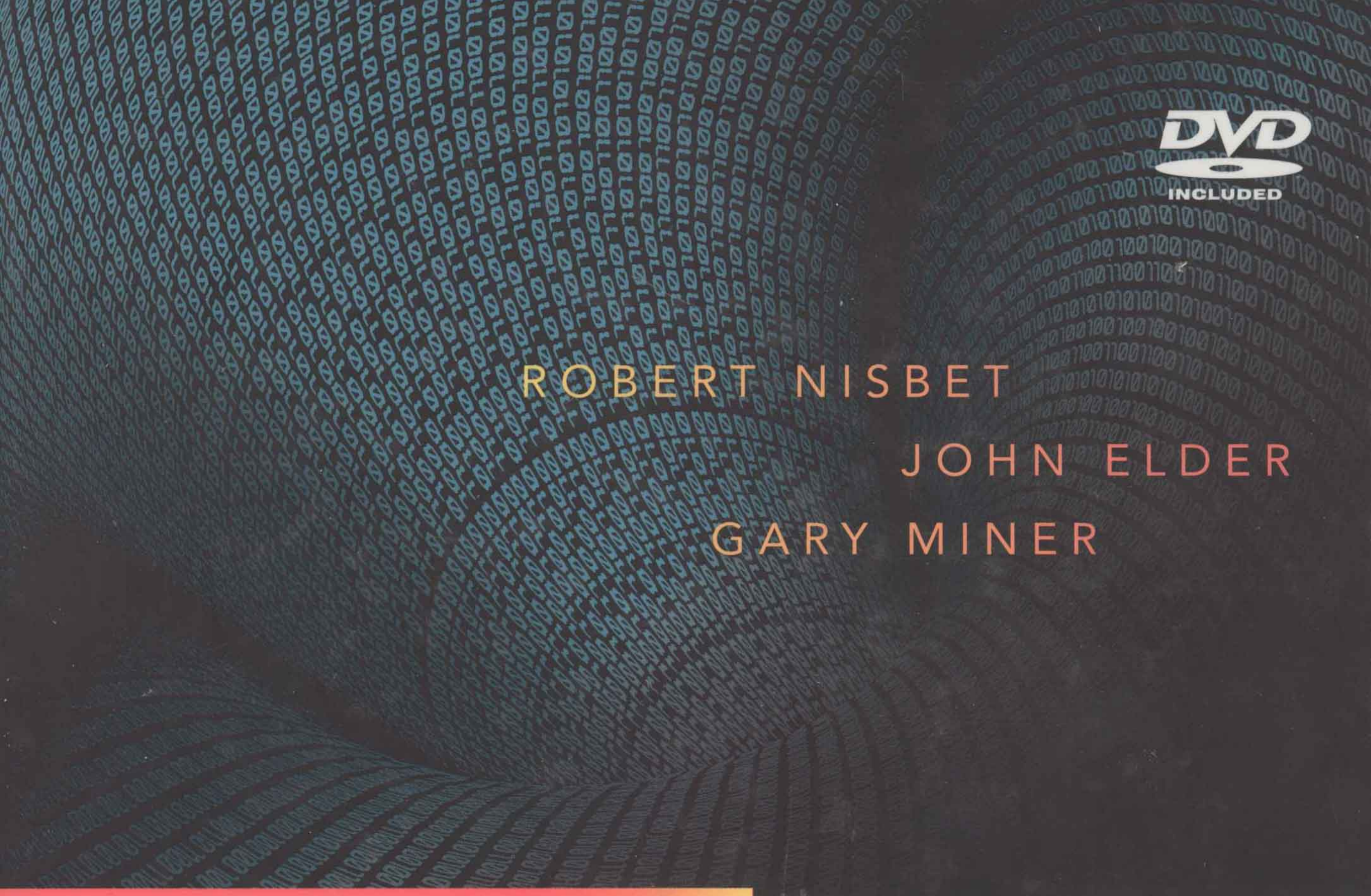

DVD INCLUDED

ROBERT NISBET

JOHN ELDER

GARY MINER

HANDBOOK OF Statistical Analysis & Data Mining Applications

HANDBOOK OF STATISTICAL ANALYSIS AND DATA MINING APPLICATIONS

江苏工业学院图书馆
藏书章

ROBERT NISBET
Pacific Capital Bankcorp N.A.
Santa Barbara, CA

JOHN ELDER
Elder Research, Inc., Charlottesville, VA

GARY MINER
StatSoft, Inc., Tulsa, Oklahoma

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

Copyright © 2009, Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, E-mail: permissions@elsevier.com. You may also complete your request online via the Elsevier homepage (http://elsevier.com), by selecting "Support & Contact" then "Copyright and Permission" and then "Obtaining Permissions."

Library of Congress Cataloging-in-Publication Data
Nisber, Robert, 1942-
Handbook of statistical analysis and data mining applications / Robert Nisbet, John Elder, Gary Miner.
p. cm.
Includes index.
ISBN 978-0-12-374765-5 (hardcover : alk. pager) 1. Data mining–Statistical methods. I. Elder, John F. (John Fletcher) II. Miner, Gary. III. Title.
QA76.9.D343N57 2009
006.3′12–dc22 2009008997

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

ISBN: 978-0-12-374765-5

For information on all Academic Press publications
visit our Web site at www.elsevierdirect.com

Printed in Canada
09 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

Table of Contents

Foreword 1 xv
Foreword 2 xvii
Preface xix
Introduction xxiii
List of Tutorials by Guest Authors xxix

I

HISTORY OF PHASES OF DATA ANALYSIS, BASIC THEORY, AND THE DATA MINING PROCESS

1. The Background for Data Mining Practice

Preamble 3
A Short History of Statistics and Data Mining 4
Modern Statistics: A Duality? 5
Assumptions of the Parametric Model 6
Two Views of Reality 8
Aristotle 8
Plato 9
The Rise of Modern Statistical Analysis: The Second Generation 10
Data, Data Everywhere . . . 11
Machine Learning Methods: The Third Generation 11
Statistical Learning Theory: The Fourth Generation 12
Postscript 13

2. Theoretical Considerations for Data Mining

Preamble 15
The Scientific Method 16
What Is Data Mining? 17
A Theoretical Framework for the Data Mining Process 18
Microeconomic Approach 19
Inductive Database Approach 19
Strengths of the Data Mining Process 19
Customer-Centric Versus Account-Centric: A New Way to Look at Your Data 20
The Physical Data Mart 20
The Virtual Data Mart 21
Householded Databases 21
The Data Paradigm Shift 22
Creation of the Car 22
Major Activities of Data Mining 23
Major Challenges of Data Mining 25
Examples of Data Mining Applications 26
Major Issues in Data Mining 26
General Requirements for Success in a Data Mining Project 28
Example of a Data Mining Project: Classify a Bat's Species by Its Sound 28
The Importance of Domain Knowledge 30
Postscript 30
Why Did Data Mining Arise? 30
Some Caveats with Data Mining Solutions 31

3. The Data Mining Process

Preamble 33
The Science of Data Mining 33
The Approach to Understanding and Problem Solving 34
CRISP-DM 35
Business Understanding (Mostly Art) 36
Define the Business Objectives of the Data Mining Model 36
Assess the Business Environment for Data Mining 37
Formulate the Data Mining Goals and Objectives 37

- Data Understanding (Mostly Science) 39
  - Data Acquisition 39
  - Data Integration 39
  - Data Description 40
  - Data Quality Assessment 40
- Data Preparation (A Mixture of Art and Science) 40
- Modeling (A Mixture of Art and Science) 41
  - Steps in the Modeling Phase of CRISP-DM 41
- Deployment (Mostly Art) 45
- Closing the Information Loop (Art) 46
- The Art of Data Mining 46
  - Artistic Steps in Data Mining 47
- Postscript 47

4. Data Understanding and Preparation

- Preamble 49
- Activities of Data Understanding and Preparation 50
  - Definitions 50
- Issues That Should be Resolved 51
  - Basic Issues That Must Be Resolved in Data Understanding 51
  - Basic Issues That Must Be Resolved in Data Preparation 51
- Data Understanding 51
  - Data Acquisition 51
  - Data Extraction 53
  - Data Description 54
  - Data Assessment 56
  - Data Profiling 56
  - Data Cleansing 56
  - Data Transformation 57
  - Data Imputation 59
  - Data Weighting and Balancing 62
  - Data Filtering and Smoothing 64
  - Data Abstraction 66
  - Data Reduction 69
  - Data Sampling 69
  - Data Discretization 73
  - Data Derivation 73
- Postscript 75

5. Feature Selection

- Preamble 77
- Variables as Features 78
- Types of Feature Selections 78
- Feature Ranking Methods 78
  - Gini Index 78
  - Bi-variate Methods 80
  - Multivariate Methods 80
  - Complex Methods 82
- Subset Selection Methods 82
  - The Other Two Ways of Using Feature Selection in *STATISTICA*: Interactive Workspace 93
  - *STATISTICA* DMRecipe Method 93
- Postscript 96

6. Accessory Tools for Doing Data Mining

- Preamble 99
- Data Access Tools 100
  - Structured Query Language (SQL) Tools 100
  - Extract, Transform, and Load (ETL) Capabilities 100
- Data Exploration Tools 101
  - Basic Descriptive Statistics 101
  - Combining Groups (Classes) for Predictive Data Mining 105
  - Slicing/Dicing and Drilling Down into Data Sets/ Results Spreadsheets 106
- Modeling Management Tools 107
  - Data Miner Workspace Templates 107
- Modeling Analysis Tools 107
  - Feature Selection 107
  - Importance Plots of Variables 108
- In-Place Data Processing (IDP) 113
  - Example: The IDP Facility of *STATISTICA* Data Miner 114
  - How to Use the SQL 114
- Rapid Deployment of Predictive Models 114
- Model Monitors 116
- Postscript 117

试读结束，需要全本PDF请购买 www.ertongbook.com

II

THE ALGORITHMS IN DATA MINING AND TEXT MINING, THE ORGANIZATION OF THE THREE MOST COMMON DATA MINING TOOLS, AND SELECTED SPECIALIZED AREAS USING DATA MINING

7. Basic Algorithms for Data Mining: A Brief Overview

Preamble 121
- *STATISTICA* Data Miner Recipe (DMRecipe) 123
- KXEN 124

Basic Data Mining Algorithms 126
- Association Rules 126
- Neural Networks 128
- Radial Basis Function (RBF) Networks 136
- Automated Neural Nets 138

Generalized Additive Models (GAMs) 138
- Outputs of GAMs 139
- Interpreting Results of GAMs 139

Classification and Regression Trees (CART) 139
- Recursive Partitioning 144
- Pruning Trees 144
- General Comments about CART for Statisticians 144
- Advantages of CART over Other Decision Trees 145
- Uses of CART 146

General CHAID Models 146
- Advantages of CHAID 147
- Disadvantages of CHAID 147

Generalized EM and *k*-Means Cluster Analysis—An Overview 147
- *k*-Means Clustering 147
- EM Cluster Analysis 148
- Processing Steps of the EM Algorithm 149
- *V*-fold Cross-Validation as Applied to Clustering 149

Postscript 150

8. Advanced Algorithms for Data Mining

Preample 151

Advanced Data Mining Algorithms 154
- Interactive Trees 154
- Multivariate Adaptive Regression Splines (MARSplines) 158
- Statistical Learning Theory: Support Vector Machines 162
- Sequence, Association, and Link Analyses 164
- Independent Components Analysis (ICA) 168
- Kohonen Networks 169
- Characteristics of a Kohonen Network 169
- Quality Control Data Mining and Root Cause Analysis 169

Image and Object Data Mining: Visualization and 3D-Medical and Other Scanning Imaging 170

Postscript 171

9. Text Mining and Natural Language Processing

Preamble 173

The Development of Text Mining 174

A Practical Example: NTSB 175
- Goals of Text Mining of NTSB Accident Reports 184
- Drilling into Words of Interest 188
- Means with Error Plots 189
- Feature Selection Tool 190
- A Conclusion: Losing Control of the Aircraft in Bad Weather Is Often Fatal 191
- Summary 194

Text Mining Concepts Used in Conducting Text Mining Studies 194

Postscript 194

10. The Three Most Common Data Mining Software Tools

Preamble 197

SPSS Clementine Overview 197
- Overall Organization of Clementine Components 198
- Organization of the Clementine Interface 199
- Clementine Interface Overview 199
- Setting the Default Directory 201
- SuperNodes 201

Execution of Streams 202
SAS-Enterprise Miner (SAS-EM) Overview 203
Overall Organization of SAS-EM Version 5.3 Components 203
Layout of the SAS-Enterprise Miner Window 204
Various SAS-EM Menus, Dialogs, and Windows Useful During the Data Mining Process 205
Software Requirements to Run SAS-EM 5.3 Software 206
STATISTICA Data Miner, QC-Miner, and Text Miner Overview 214
Overall Organization and Use of *STATISTICA* Data Miner 214
Three Formats for Doing Data Mining in *STATISTICA* 230
Postscript 234

11. Classification

Preample 235
What Is Classification? 235
Initial Operations in Classification 236
Major Issues with Classification 236
What Is the Nature of Data Set to Be Classified? 236
How Accurate Does the Classification Have to Be? 236
How Understandable Do the Classes Have to Be? 237
Assumptions of Classification Procedures 237
Numerical Variables Operate Best 237
No Missing Values 237
Variables Are Linear and Independent in Their Effects on the Target Variable 237
Methods for Classification 238
Nearest-Neighbor Classifiers 239
Analyzing Imbalanced Data Sets with Machine Learning Programs 240
CHAID 246
Random Forests and Boosted Trees 248
Logistic Regression 250
Neural Networks 251
Naïve Bayesian Classifiers 253
What Is the Best Algorithm for Classification? 256
Postscript 257

12. Numerical Prediction

Preamble 259
Linear Response Analysis and the Assumptions of the Parametric Model 260
Parametric Statistical Analysis 261
Assumptions of the Parametric Model 262
The Assumption of Independency 262
The Assumption of Normality 262
Normality and the Central Limit Theorem 263
The Assumption of Linearity 264
Linear Regression 264
Methods for Handling Variable Interactions in Linear Regression 265
Collinearity among Variables in a Linear Regression 265
The Concept of the Response Surface 266
Generalized Linear Models (GLMs) 270
Methods for Analyzing Nonlinear Relationships 271
Nonlinear Regression and Estimation 271
Logit and Probit Regression 272
Poisson Regression 272
Exponential Distributions 272
Piecewise Linear Regression 273
Data Mining and Machine Learning Algorithms Used in Numerical Prediction 274
Numerical Prediction with C&RT 274
Model Results Available in C&RT 276
Advantages of Classification and Regression Trees (C&RT) Methods 277
General Issues Related to C&RT 279
Application to Mixed Models 280
Neural Nets for Prediction 280
Manual or Automated Operation? 280
Structuring the Network for Manual Operation 280
Modern Neural Nets Are "Gray Boxes" 281
Example of Automated Neural Net Results 281
Support Vector Machines (SVMs) and Other Kernel Learning Algorithms 282
Postscript 284

13. Model Evaluation and Enhancement

Preamble 285
Introduction 286
Model Evaluation 286
Splitting Data 287

Avoiding Overfit Through Complexity Regularization 288
Error Metric: Estimation 291
Error Metric: Classification 291
Error Metric: Ranking 293
Cross-Validation to Estimate Error Rate and Its Confidence 295
Bootstrap 296
Target Shuffling to Estimate Baseline Performance 297
Re-Cap of the Most Popular Algorithms 300
Linear Methods (Consensus Method, Stepwise Is Variable-Selecting) 300
Decision Trees (Consensus Method, Variable-Selecting) 300
Neural Networks (Consensus Method) 301
Nearest Neighbors (Contributory Method) 301
Clustering (Consensus or Contributory Method) 302
Enhancement Action Checklist 302
Ensembles of Models: The Single Greatest Enhancement Technique 304
Bagging 305
Boosting 305
Ensembles in General 306
How to Thrive as a Data Miner 307
Big Picture of the Project 307
Project Methodology and Deliverables 308
Professional Development 309
Three Goals 310
Postscript 311

14. Medical Informatics
Preamble 313
What Is Medical Informatics? 313
How Data Mining and Text Mining Relate to Medical Informatics 314
XplorMed 316
ABView: HivResist 317
3D Medical Informatics 317
What Is 3D Informatics? 317
Future and Challenges of 3D Medical Informatics 318
Journals and Associations in the Field of Medical Informatics 318
Postscript 318

15. Bioinformatics
Preamble 321
What Is Bioinformatics? 323
Data Analysis Methods in Bioinformatics 326
ClustalW2: Sequence Alignment 326
Searching Databases for RNA Molecules 327
Web Services in Bioinformatics 327
How Do We Apply Data Mining Methods to Bioinformatics? 329
Postscript 332
Tutorial Associated with This Chapter on Bioinformatics 332
Books, Associations, and Journals on Bioinformatics, and Other Resources, Including Online 332

16. Customer Response Modeling
Preamble 335
Early CRM Issues in Business 336
Knowing How Customers Behaved Before They Acted 336
Transforming Corporations into Business Ecosystems: The Path to Customer Fulfillment 337
CRM in Business Ecosystems 338
Differences Between Static Measures and Evolutionary Measures 338
How Can Human Nature as Viewed Through Plato Help Us in Modeling Customer Response? 339
How Can We Reorganize Our Data to Reflect Motives and Attitudes? 339
What Is a Temporal Abstraction? 340
Conclusions 344
Postscript 345

17. Fraud Detection
Preamble 347
Issues with Fraud Detection 348
Fraud Is Rare 348
Fraud Is Evolving! 348
Large Data Sets Are Needed 348
The Fact of Fraud Is Not Always Known during Modeling 348
When the Fraud Happened Is Very Important to Its Detection 349

Fraud Is Very Complex 349
Fraud Detection May Require the Formulation of Rules Based on General Principles,"Red Flags," Alerts, and Profiles 349
Fraud Detection Requires Both Internal and External Business Data 349
Very Few Data Sets and Modeling Details Are Available 350
How Do You Detect Fraud? 350
Supervised Classification of Fraud 351
How Do You Model Fraud? 352
How Are Fraud Detection Systems Built? 353
Intrusion Detection Modeling 355
Comparison of Models with and without Time-Based Features 355
Building Profiles 360
Deployment of Fraud Profiles 360
Postscript and Prolegomenon 361

III

TUTORIALS—STEP-BY-STEP CASE STUDIES AS A STARTING POINT TO LEARN HOW TO DO DATA MINING ANALYSES

Guest Authors of the Tutorials

A. How to Use Data Miner Recipe
What is *STATISTICA* Data Miner Recipe (DMR)? 373
Core Analytic Ingredients 373

B. Data Mining for Aviation Safety
Airline Safety 378
SDR Database 379
Preparing the Data for Our Tutorial 382
Data Mining Approach 383
Data Mining Algorithm Error Rate 386
Conclusion 387

C. Predicting Movie Box-Office Receipts
Introduction 391
Data and Variable Definitions 392
Getting to Know the Workspace of the Clementine Data Mining Toolkit 393
Results 396
Publishing and Reuse of Models and Other Outputs 404

D. Detecting Unsatisfied Customers: A Case Study
Introduction 418
The Data 418
The Objectives of the Study 418
SAS-EM 5.3 Interface 419
A Primer of SAS-EM Predictive Modeling 420
Homework 1 430
Discussions 431
Homework 2 431
Homework 3 431
Scoring Process and the Total Profit 432
Homework 4 438
Discussions 439
Oversampling and Rare Event Detection 439
Discussion 446
Decision Matrix and the Profit Charts 446
Discussions 453
Micro-Target the Profitable Customers 453
Appendix 455

E. Credit Scoring
Introduction: What Is Credit Scoring? 459
Credit Scoring: Business Objectives 460
Case Study: Consumer Credit Scoring 461
Description 461
Data Preparation 462
Feature Selection 462
STATISTICA Data Miner: "Workhorses" or Predictive Modeling 463
Overview: *STATISTICA* Data Miner Workspace 464
Analysis and Results 465
Decision Tree: CHAID 465
Classification Matrix: CHAID Model 467

Comparative Assessment of the Models (Evaluation) 467
Classification Matrix: Boosting Trees with Deployment Model (Best Model) 469
Deploying the Model for Prediction 469
Conclusion 470

F. Churn Analysis
Objectives 471
Steps 472

G. Text Mining: Automobile Brand Review
Introduction 481
Text Mining 482
Input Documents 482
Selecting Input Documents 482
Stop Lists, Synonyms, and Phrases 482
Stemming and Support for Different Languages 483
Indexing of Input Documents: Scalability of *STATISTICA* Text Mining and Document Retrieval 483
Results, Summaries, and Transformations 483
Car Review Example 484
Saving Results into Input Spreadsheet 498
Interactive Trees (C&RT, CHAID) 503
Other Applications of Text Mining 512
Conclusion 512

H. Predictive Process Control: QC-Data Mining
Predictive Process Control Using *STATISTICA* and *STATISTICA* Qc-miner 513
Case Study: Predictive Process Control 514
Understanding Manufacturing Processes 514
Data File: ProcessControl.sta 515
Variable Information 515
Problem Definition 515
Design Approaches 515
Data Analyses with *STATISTICA* 517
Split Input Data into the Training and Testing Sample 517
Stratified Random Sampling 517
Feature Selection and Root Cause Analyses 517
Different Models Used for Prediction 518
Compute Overlaid Lift Charts from All Models: Static Analyses 520
Classification Trees: CHAID 521
Compute Overlaid Lift/Gain Charts from All Models: Dynamic Analyses 523
Cross-Tabulation Matrix 524
Comparative Evaluation of Models: Dynamic Analyses 526
Gains Analyses by Deciles: Dynamic Analyses 526
Transformation of Change 527
Feature Selection and Root Cause Analyses 528
Interactive Trees: C&RT 528
Conclusion 529

I. Business Administration in a Medical Industry

J. Clinical Psychology: Making Decisions about Best Therapy for a Client

K. Education–Leadership Training for Business and Education

L. Dentistry: Facial Pain Study

M. Profit Analysis of the German Credit Data
Introduction 651
Modeling Strategy 653
SAS-EM 5.3 Interface 654
A Primer of SAS-EM Predictive Modeling 654
Advanced Techniques of Predictive Modeling 669
Micro-Target the Profitable Customers 676
Appendix 678

N. Predicting Self-Reported Health Status Using Artificial Neural Networks
Background 681
Data 682
Preprocessing and Filtering 683

Part 1: Using a Wrapper Approach in Weka to Determine the Most Appropriate Variables for Your Neural Network Model 684
Part 2: Taking the Results from the Wrapper Approach in Weka into *STATISTICA* Data Miner to Do Neural Network Analyses 691

IV

MEASURING TRUE COMPLEXITY, THE "RIGHT MODEL FOR THE RIGHT USE," TOP MISTAKES, AND THE FUTURE OF ANALYTICS

18. Model Complexity (and How Ensembles Help)
Preamble 707
Model Ensembles 708
Complexity 710
Generalized Degrees of Freedom 713
Examples: Decision Tree Surface with Noise 714
Summary and Discussion 719
Postscript 720

19. The Right Model for the Right Purpose: When Less Is Good Enough
Preamble 723
More Is not Necessarily Better: Lessons from Nature and Engineering 724
Embrace Change Rather Than Flee from It 725
Decision Making Breeds True in the Business Organism 725
Muscles in the Business Organism 726
What Is a Complex System? 726
The 80:20 Rule in Action 728
Agile Modeling: An Example of How to Craft Sufficient Solutions 728
Postscript 730

20. Top 10 Data Mining Mistakes
Preamble 733
Introduction 734
0. Lack Data 734
1. Focus on Training 735
2. Rely on One Technique 736
3. Ask the Wrong Question 738
4. Listen (Only) to the Data 739
5. Accept Leaks from the Future 742
6. Discount Pesky Cases 743
7. Extrapolate 744
8. Answer Every Inquiry 747
9. Sample Casually 750
10. Believe the Best Model 752
How Shall We Then Succeed? 753
Postscript 753

21. Prospects for the Future of Data Mining and Text Mining as Part of Our Everyday Lives
Preamble 755
RFID 756
Social Networking and Data Mining 757
Example 1 758
Example 2 759
Example 3 760
Example 4 761
Image and Object Data Mining 761
Visual Data Preparation for Data Mining: Taking Photos, Moving Pictures, and Objects into Spreadsheets Representing the Photos, Moving Pictures, and Objects 765
Cloud Computing 769
What Can Science Learn from Google? 772
The Next Generation of Data Mining 772
From the Desktop to the Clouds . . . 778
Postscript 778

22. Summary: Our Design
Preamble 781
Beware of Overtrained Models 782
A Diversity of Models and Techniques Is Best 783
The Process Is More Important Than the Tool 783

试读结束，需要全本PDF请购买 www.ertongbook.com

Text Mining of Unstructured Data Is Becoming Very Important 784
Practice Thinking About Your Organization as *Organism* Rather Than as *Machine* 784
Good Solutions Evolve Rather Than Just Appear After Initial Efforts 785
What You *Don't* Do Is Just as Important as What You *Do* 785
Very Intuitive Graphical Interfaces Are Replacing Procedural Programming 786
Data Mining Is No Longer a Boutique Operation; It Is Firmly Established in the Mainstream of Our Society 786
"Smart" Systems Are the Direction in Which Data Mining Technology Is Going 787
Postscript 787

Glossary 789

Index 801

DVD Install Instructions 823

Foreword 1

This book will help the novice user become familiar with data mining. Basically, data mining is doing data analysis (or statistics) on data sets (often large) that have been obtained from potentially many sources. As such, the miner may not have control of the input data, but must rely on sources that have gathered the data. As such, there are problems that every data miner must be aware of as he or she begins (or completes) a mining operation. I strongly resonated to the material on "The Top 10 Data Mining Mistakes," which give a worthwhile checklist:

- Ensure you have a response variable and predictor variables—and that they are correctly measured.
- Beware of overfitting. With scads of variables, it is easy with most statistical programs to fit incredibly complex models, but they cannot be reproduced. It is good to save part of the sample to use to test the model. Various methods are offered in this book.
- Don't use only one method. Using only linear regression can be a problem. Try dichotomizing the response or categorizing it to remove nonlinearities in the response variable. Often, there are clusters of values at zero, which messes up any normality assumption. This, of course, loses information, so you may want to categorize a continuous response variable and use an alternative to regression. Similarly, predictor variables may need to be treated as factors rather than linear predictors. A classic example is using marital status or race as a linear predictor when there is no order.
- Asking the wrong question—when looking for a rare phenomenon, it may be helpful to identify the most common pattern. These may lead to complex analyses, as in item 3, but they may also be conceptually simple. Again, you may need to take care that you don't overfit the data.
- Don't become enamored with the data. There may be a substantial history from earlier data or from domain experts that can help with the modeling.
- Be wary of using an outcome variable (or one highly correlated with the outcome variable) and becoming excited about the result. The predictors should be "proper" predictors in the sense that (a) they are measured prior to the outcome and (b) are not a function of the outcome.
- Do not discard outliers without solid justification. Just because an observation is out of line with others is insufficient reason to ignore it. You must check the circumstances that led to the value. In any event, it is useful to conduct the analysis with the observation(s) included and excluded to determine the sensitivity of the results to the outlier.

- Extrapolating is a fine way to go broke—the best example is the stock market. Stick within your data, and if you must go outside, put plenty of caveats. Better still, restrain the impulse to extrapolate. Beware that pictures are often far too simple and we can be misled. Political campaigns oversimplify complex problems ("My opponent wants to raise taxes"; "My opponent will take us to war") when the realities may imply we have some infrastructure needs that can be handled only with new funding, or we have been attacked by some bad guys.

Be wary of your data sources. If you are combining several sets of data, they need to meet a few standards:

- The definitions of variables that are being merged should be identical. Often they are close but not exact (especially in meta-analysis where clinical studies may have somewhat different definitions due to different medical institutions or laboratories).
- Be careful about missing values. Often when multiple data sets are merged, missing values can be induced: one variable isn't present in another data set, what you thought was a unique variable name was slightly different in the two sets, so you end up with two variables that both have a lot of missing values.
- How you handle missing values can be crucial. In one example, I used complete cases and lost half of my sample—all variables had at least 85% completeness, but when put together the sample lost half of the data. The residual sum of squares from a stepwise regression was about 8. When I included more variables using mean replacement, almost the same set of predictor variables surfaced, but the residual sum of squares was 20. I then used multiple imputation and found approximately the same set of predictors but had a residual sum of squares (median of 20 imputations) of 25. I find that mean replacement is rather optimistic but surely better than relying on only complete cases. If using stepwise regression, I find it useful to replicate it with a bootstrap or with multiple imputation. However, with large data sets, this approach may be expensive computationally.

To conclude, there is a wealth of material in this handbook that will repay study.

Peter A. Lachenbruch, Ph.D.,
Oregon State University
Past President, 2008, American Statistical Society
Professor, Oregon State University
Formerly: FDA and professor at Johns Hopkins University;
UCLA, and University of Iowa, and
University of North Carolina Chapel Hill

Foreword 2

A November 2008 search on Amazon.com for "data mining" books yielded over 15,000 hits—including 72 to be published in 2009. Most of these books either describe data mining in very technical and mathematical terms, beyond the reach of most individuals, or approach data mining at an introductory level without sufficient detail to be useful to the practitioner. The *Handbook of Statistical Analysis and Data Mining Applications* is the book that strikes the right balance between these two treatments of data mining.

This volume is not a theoretical treatment of the subject—the authors themselves recommend other books for this—but rather contains a description of data mining principles and techniques in a series of "knowledge-transfer" sessions, where examples from real data mining projects illustrate the main ideas. This aspect of the book makes it most valuable for practitioners, whether novice or more experienced.

While it would be easier for everyone if data mining were merely a matter of finding and applying the correct mathematical equation or approach for any given problem, the reality is that both "art" and "science" are necessary. The "art" in data mining requires experience: when one has seen and overcome the difficulties in finding solutions from among the many possible approaches, one can apply newfound wisdom to the next project. However, this process takes considerable time and, particularly for data mining novices, the iterative process inevitable in data mining can lead to discouragement when a "textbook" approach doesn't yield a good solution.

This book is different; it is organized with the practitioner in mind. The volume is divided into four parts. Part I provides an overview of analytics from a historical perspective and frameworks from which to approach data mining, including CRISP-DM and SEMMA. These chapters will provide a novice analyst an excellent overview by defining terms and methods to use, and will provide program managers a framework from which to approach a wide variety of data mining problems. Part II describes algorithms, though without extensive mathematics. These will appeal to practitioners who are or will be involved with day-to-day analytics and need to understand the qualitative aspects of the algorithms. The inclusion of a chapter on text mining is particularly timely, as text mining has shown tremendous growth in recent years.

Part III provides a series of tutorials that are both domain-specific and software-specific. Any instructor knows that examples make the abstract concept more concrete, and these tutorials accomplish exactly that. In addition, each tutorial shows how the solutions were developed using popular data mining software tools, such as Clementine, Enterprise Miner, Weka, and *STATISTICA*. The step-by-step specifics will assist practitioners in learning not only how to approach a wide variety of problems, but also how to use these software products effectively. Part IV presents a look at the future of data mining, including a treatment of model ensembles and "The Top 10 Data Mining Mistakes," from the popular presentation by Dr. Elder.

However, the book is best read a few chapters at a time while actively doing the data mining rather than read cover-to-cover (a daunting task for a book this size). Practitioners will appreciate tutorials that match their business objectives and choose to ignore other tutorials. They may choose to read sections on a particular algorithm to increase insight into that algorithm and then decide to add a second algorithm after the first is mastered. For those new to a particular software tool highlighted in the tutorials section, the step-by-step approach will operate much like a user's manual. Many chapters stand well on their own, such as the excellent "History of Statistics and Data Mining" and "The Top 10 Data Mining Mistakes" chapters. These are broadly applicable and should be read by even the most experienced data miners.

The *Handbook of Statistical Analysis and Data Mining Applications* is an exceptional book that should be on every data miner's bookshelf or, better yet, found lying open next to their computer.

Dean Abbott
President
Abbott Analytics
San Diego, California

Preface

Data mining scientists in research and academia may look askance at this book because it does not present algorithm theory in the commonly accepted mathematical form. Most articles and books on data mining and knowledge discovery are packed with equations and mathematical symbols that only experts can follow. Granted, there is a good reason for insistence on this formalism. The underlying complexity of nature and human response requires teachers and researchers to be extremely clear and unambiguous in their terminology and definitions. Otherwise, ambiguities will be communicated to students and readers, and their understanding will not penetrate to the essential elements of any topic. Academic areas of study are not called *disciplines* without reason.

This rigorous approach to data mining and knowledge discovery builds a fine foundation for academic studies and research by experts. Excellent examples of such books are

- *The Handbook of Data Mining*, 2003, by Nong Ye (Ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, 2009, by T. Hastie, R. Tibshirani, & J. Friedman. New York: Springer-Verlag.

Books like these were especially necessary in the early days of data mining, when analytical tools were relatively crude and required much manual configuration to make them work right. Early users had to understand the tools in depth to be able to use them productively. These books are still necessary for the college classroom and research centers. Students must understand the theory behind these tools in the same way that the developers understood it so that they will be able to build new and improved versions.

Modern data mining tools, like the ones featured in this book, permit ordinary business analysts to follow a path through the data mining process to create models that are "good enough." These less-than-optimal models are far better in their ability to leverage faint patterns in databases to solve problems than the ways it used to be done. These tools provide default configurations and automatic operations, which shield the user from the technical complexity underneath. They provide one part in the crude analogy to the automobile interface. You don't have to be a chemical engineer or physicist who understands moments of force to be able to operate a car. All you have to do is learn to turn the key in the ignition, step on the gas and the brake at the right times, turn the wheel to change direction in a safe manner, and *voila*, you are an expert user of the very complex technology

试读结束，需要全本PDF请购买 www.ertongbook.com