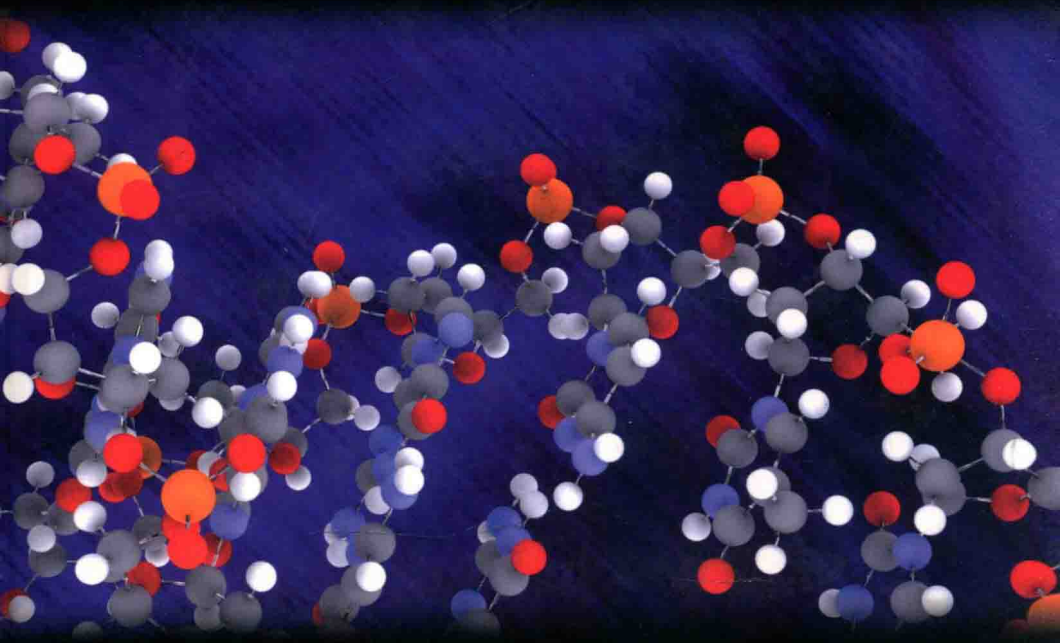
Science, Engineering, and Biology Informatics Vol. 2

life science data mining



editors

Stephen Wong • Chung-Sheng Li

life science data mining

editors

Stephen Wong
Harvard Medical School, USA

Chung-Sheng Li
IBM Thomas J Watson Research Center

NEW JERSEY · LONDON · SINGAPORE · BEIJING · SHANGHAI · HONG KONG · TAIPEI · CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

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

Science, Engineering, and Biology Informatics — Vol. 2
LIFE SCIENCE DATA MINING

ISBN 981-270-064-1
ISBN 981-270-065-X (pbk)

Printed in Singapore by B & JO Enterprise

life science data mining

SCIENCE, ENGINEERING, AND BIOLOGY INFORMATICS

Series Editor: Jason T. L. Wang
(*New Jersey Institute of Technology, USA*)

Published:

Vol. 1: Advanced Analysis of Gene Expression Microarray Data
(*Aidong Zhang*)

Vol. 2: Life Science Data Mining
(*Stephen TC Wong & Chung-Sheng Li*)

Forthcoming:

Vol. 3: Analysis of Biological Data: A Soft Computing Approach
(*Sanghamitra Bandyopadhyay, Ujjwal Maulik & Jason T. L. Wang*)

PREFACE

Data mining is the process of using computational algorithms and tools to automatically discover useful information in large data archives. Data mining techniques are deployed to score large databases in order to find novel and useful patterns that might otherwise remain unknown. They also can be used to predict the outcome of a future observation or to assess the potential risk in a disease situation. Recent advances in data generation devices, data acquisition, and storage technology in the life sciences have enabled biomedical research and healthcare organizations to accumulate vast amounts of heterogeneous data that is key to important new discoveries or therapeutic interventions. Extracting useful information has proven extremely challenging however. Traditional data analysis and mining tools and techniques often cannot be used because of the massive size of a data set and the non-traditional nature of the biomedical data, compared to those encountered in financial and commercial sectors. In many situations, the questions that need to be answered cannot be addressed using existing data analysis and mining techniques, and thus, new algorithms and methods need to be developed.

Life science is an important application domain that requires new techniques of data analysis and mining. This is one of the first technical books focusing on the data analysis and mining techniques in life science applications. In this introductory chapter, we present the key topics to be covered in this book. In Chapter 1 "Taxonomy of early detection for environmental and public health applications," Chung-Sheng Li of IBM Research provides a survey of early warning systems and detection approaches in terms of problem domains and data sources. The chapter introduces current syndromic surveillance prototypes or deployments and defines the problem domain for three classes: individual and public health level, cellular level, and molecular level. For data sources, they were also categorized into three parts including clinically related data, non-traditional data, and auxiliary data. Furthermore, data sources can be characterized by three dimensions (structured, semi-structured, and non-structured).

In Chapter 2 "Time-lapse cell cycle quantitative data analysis using Gaussian mixture models," Xiaobo Zhou and colleagues at Harvard Medical School describe an interesting and important emerging technology area of high throughput biological imaging. The authors address the unresolved problem of identifying the cell cycle process under different conditions of perturbation. In the study, the time-lapse fluorescence microscopy imaging images are analyzed to detect and measure the duration of various cell phases, e.g., inter phase, prophase, metaphase, and anaphase, quantitatively.

Chapter 3 "Diversity and accuracy of data mining ensemble" by Wenjia Wang of the University of East Anglia discusses an important issue: classifier fusion or an ensemble of classifiers. This paper first describes why diversity is essential and how the diversity can be measured, then it analyses the relationships between the accuracy of an ensemble and the diversity among its member classifiers. An example is given to show that the mixed ensembles are able to improve the performance.

Various clustering algorithms have been applied to gene expression data analysis. Chapter 4 "Integrated clustering for microarray data" by Gabriela Moise and Jorg Sander from the University of Alberta argue that the integration strategy is a "majority voting" approach based on the assumption that objects belong to a "natural" cluster are likely to be co-located in the same cluster by different clustering algorithms. The chapter also provides an excellent survey of clustering and integrated clustering approaches in microarray analysis.

EEG has a variety of applications in basic and clinic neuroscience. Chapter 5: "Complexity and Synchronization of EEG with Parametric Modeling" by Xiaoli Li presents a nice work on parametric modeling of the complexity and synchronization of EEG signals to assist the diagnosis of epilepsy or the analysis of EEG dynamics.

In Chapter 6: "Bayesian Fusion of Syndromic Surveillance with Sensor Data for Disease Outbreak Classification," by Jeffrey Lin, Howard Burkom, *et al.* from The Johns Hopkins University Applied Physics Laboratory and Walter Reed Army Institute for Research describe a novel Bayesian approach to fuse sensor data with syndromic surveillance data presented for timely detection and classification of

disease outbreaks. In addition, the authors select a natural disease, asthma, which is highly dependent on environmental factors to validate the approach.

Continuing the theme of syndromic surveillance, in Chapter 7: "An Evaluation of Over-the-Counter Medication Sales for Syndromic Surveillance," Murray Campbell, Chung Sheng Li, *et al.* from IBM Watson Research Center describe a number of approaches to evaluate the utility of data sources in a syndromic surveillance context and show that there may be some values in using sales of over-the-counter medications for syndromic surveillance.

In Chapter 8: "Collaborative Health Sentinel" JH Kaufman, G Decad, *et al.* from IBM Research Divisions in California, New York, and Israel provide a clear survey and highlight the approach of significant trends, issues and further directions for global systems for health management. They addressed various issues for systems covering general environment and public health, as well as global view.

In Chapter 9: "Data Mining for Drug Abuse Research and Treatment Evaluation: Data Systems Needs and Challenges" Mary Lynn Brecht of UCLA argues that drug abuse research and evaluation can benefit from data mining strategies to generate models of complex dynamic phenomena from heterogeneous data sources. She presents a user's perspective and several challenges on selected topics relating to the development of an online processing framework for data retrieval and mining to meet the needs in this field.

The increasing amount and complexity of data used in predictive toxicology call for new and flexible approaches based on hybrid intelligent methods to mine the data. To fill this needs, in Chapter 10 "Knowledge Representation for Versatile Hybrid Intelligent Processing Applied in Predictive Toxicology", Daneil Neagu from Bradford University, England addresses the issue of devising a mark-up language, as an application of XML (Extensible Markup Language), for representing knowledge modeling in predictive toxicology - PToXML and the markup language HISML for integrated data structures of Hybrid Intelligent Systems.

Ensemble classification is an active field of research in pattern recognition. In Chapter 11: "Ensemble Classification System

Implementation for Biomedical Microarray Data," Shun Bian and Wenjia Wang of the University of East Anglia, UK present a framework of developing a flexible software platform for building an ensemble based on the diversity measures. An ensemble classification system (ECS) has been implemented for mining biomedical data as well as general data.

Time-lapse fluorescence microscopy imaging provides an important high throughput method to study the dynamic cell cycle process under different conditions of perturbation. The bottleneck, however, lies in the analysis and modeling of large amounts of image data generated. Chapter 12: "An Automated Method for Cell Phase Identification in High Throughput Time-Lapse Screens" by Xiaowei Chen and colleagues from Harvard Medical School describe the application of statistical and machine learning techniques to the problem of tracking and identifying the phase of individual cells in populations as a function of time using high throughput imaging techniques.

Modeling gene regulatory networks has been an active area of research in computational biology and systems biology. An important step in constructing these networks involves finding genes that have the strongest influence on the target gene. In Chapter 13, "Inference of Transcriptional Regulatory Networks based on Cancer Microarray Data," Xiaobo Zhou and Stephen Wong from Harvard Center for Neurodegeneration and Repair address this problem. They start with certain existing subnetworks and methods of transcriptional regulatory network construction and then present their new approach.

In Chapter 14: "Data Mining in Biomedicine," Lucila Ohno-Machado and Staal Vinterbo of Brigham and Women's Hospital provide an overview of data mining techniques in biomedicine. They refer to data mining as any data processing algorithm that aims to determine patterns or regularities in the data. The patterns may be used for diagnostic or prognostic purposes and the models that result from pattern recognition algorithms will be refereed to as *predictive models*, regardless of whether they are used to classify. The chapter provides readers an excellent introduction about data mining and its applications to the readers.

Association rules mining is a popular technique for the analysis of gene expression profiles like microarray data. In Chapter 15: "Mining

Multilevel Association Rules from Gene Ontology and Microarray Data," VS Tseng and SC Yang from National Cheng Kung University, Taiwan, aim at combining microarray data and existing biological network to produce multilevel association rules. They propose a new algorithm for mining gene expression transactions based on existing algorithm ML_T1LA in the context of Gene Ontology and a filter version CMAGO.

Optical biosensors are now utilized in a wide range of applications, from biological-warfare-agent detection to improving clinical diagnosis. In Chapter 16, "A Proposed Sensor-Configuration and Sensitivity Analysis of Parameters with Applications to Biosensors," HJ Halim from Liverpool JM University, England, introduces a configuration of sensor system and analytical model equations to mitigate the effects of internal and external parameter fluctuations.

The subject of data mining in life science, while relatively young compared to data mining in other application fields, such as finance and marketing, or to statistics or machine learning, is already too large to cover in a single book volume. We hope that this edition would provide the readers some of the specific challenges that motivate the development of new data mining techniques and tools in life sciences and serve as an introductory material to the researchers and practitioners interested in this exciting field of application.

Stephen TC Wong and Chung-Sheng Li

CONTENTS

Preface v

Chapter 1 Survey of Early Warning Systems for Environmental and Public Health Applications 1
1. Introduction 1
2. Disease Surveillance 3
3. Reference Architecture for Model Extraction 5
4. Problem Domain 9
5. Data Sources 10
6. Detection Methods 12
7. Summary and Conclusion 13
References 14

Chapter 2 Time-Lapse Cell Cycle Quantitative Data Analysis Using Gaussian Mixture Models 17
1. Introduction 18
2. Material and Feature Extraction 20
2.1. Material and cell feature extraction 20
2.2. Model the time-lapse data using AR model 23
3. Problem Statement and Formulation 24
4. Classification Methods 26
4.1. Gaussian mixture models and the EM algorithm 26
4.2. K-Nearest Neighbor (KNN) classifier 28
4.3. Neural networks 28
4.4. Decision tree 29
4.5. Fisher clustering 30
5. Experimental Results 30
5.1. Trace identification 31
5.2. Cell morphologic similarity analysis 33
5.3. Phase identification 35
5.4. Cluster analysis of time-lapse data 37

6. Conclusion 40
Appendix A 41
Appendix B 42
References 43

Chapter 3 Diversity and Accuracy of Data Mining Ensemble 47
1. Introduction 47
2. Ensemble and Diversity 49
2.1. Why needs diversity? 49
2.2. Diversity measures 51
3. Probability Analysis 52
4. Coincident Failure Diversity 52
5. Ensemble Accuracy 55
5.1. Relationship between random guess and accuracy of lower bound single models 55
5.2. Relationship between accuracy A and the number of models N 56
5.3. When model's accuracy < 50% 57
6. Construction of Effective Ensembles 58
6.1. Strategies for increasing diversity 59
6.2. Ensembles of neural networks 60
6.3. Ensembles of decision trees 61
6.4. Hybrid ensembles 62
7. An Application: Osteoporosis Classification Problem 62
7.1. Osteoporosis problem 63
7.2. Results from the ensembles of neural nets 63
7.3. Results from ensembles of the decision trees 66
7.4. Results of hybrid ensembles 67
8. Discussion and Conclusions 68
References 70

Chapter 4 Integrated Clustering for Microarray Data 73
1. Introduction 73
2. Related Work 77
3. Data Preprocessing 81

4. Integrated Clustering 83
4.1. Clustering algorithms 83
4.2. Integration methodology 88
5. Experimental Evaluation 89
5.1. Evaluation methodology 89
5.2. Results 91
5.3. Discussion 93
6. Conclusions 94
References 94

Chapter 5 Complexity and Synchronization of EEG with Parametric Modeling 99
1. Introduction 100
1.1. Brief review of EEG recording analysis 100
1.2. AR modeling based EEG analysis 101
2. TVAR Modeling 104
3. Complexity Measure 105
4. Synchronization Measure 109
5. Conclusions 113
References 114

Chapter 6 Bayesian Fusion of Syndromic Surveillance with Sensor Data for Disease Outbreak Classification 119
1. Introduction 120
2. Approach 122
2.1. Bayesian belief networks 122
2.2. Syndromic data 126
2.3. Environmental data 128
2.4. Test scenarios 130
2.5. Evaluation metrics 130
3. Results 131
3.1. Scenario 1 131
3.2. Scenario 2 134
3.3. Promptness 135
4. Summary and Conclusions 136
References 137

Chapter 7 An Evaluation of Over-the-Counter Medication Sales for Syndromic Surveillance 143
1. Introduction 143
2. Background and Related Work 144
3. Data 144
4. Approaches 145
4.1. Lead-lag correlation analysis 145
4.2. Regression test of predictive ability 146
4.3. Detection-based approaches 148
4.4. Supervised algorithm for outbreak detection in OTC data 148
4.5. Modified Holt-Winters forecaster 150
4.6. Forecasting based on multi-channel regression 151
5. Experiments 153
5.1. Lead-lag correlation analysis of OTC data 153
5.2. Regression test of the predicative value of OTC 154
5.3. Results from detection-based approaches 156
6. Conclusions and Future Work 158
References 159

Chapter 8 Collaborative Health Sentinel 163
1. Introduction 163
2. Infectious Disease and Existing Health Surveillance Programs 166
3. Elements of the Collaborative Health Sentinel (CHS) System .. 170
3.1. Sampling 170
3.2. Creating a national health map 177
3.3. Detection 177
3.4. Reaction 183
3.5. Cost considerations 184
4. Interaction with the Health Information Technology (HCIT) World 185
5. Conclusion 188
References 189
Appendix A - HL7 192

Chapter 9 A Multi-Modal System Approach for Drug Abuse Research and Treatment Evaluation: Information Systems Needs and Challenges 195
1. Introduction 195
2. Context 198
2.1. Data sources 198
2.2. Examples of relevant questions 199
3. Possible System Structure 201
4. Challenges in System Development and Implementation 204
4.1. Ontology development 204
4.2. Data source control, proprietary issues 205
4.3. Privacy, security issues 205
4.4. Costs to implement/maintain system 206
4.5. Historical hypothesis-testing paradigm 206
4.6. Utility, usability, credibility of such a system 206
4.7. Funding of system development 207
5. Summary 207
References 208

Chapter 10 Knowledge Representation for Versatile Hybrid Intelligent Processing Applied in Predictive Toxicology .. 213
1. Introduction 214
2. Hybrid Intelligent Techniques for Predictive Toxicology Knowledge Representation 217
3. XML Schemas for Knowledge Representation and Processing in AI and Predictive Toxicology 218
4. Towards a Standard for Chemical Data Representation in Predictive Toxicology 220
5. Hybrid Intelligent Systems for Knowledge Representation in Predictive Toxicology 225
5.1. A formal description of implicit and explicit knowledge-based intelligent systems 226
5.2. An XML schema for hybrid intelligent systems 228
6. A Case Study 231
6.1. Materials and methods 232
6.2. Results 233

7. Conclusions 235
References 236

Chapter 11 Ensemble Classification System Implementation for Biomedical Microarray Data 239
1. Introduction 240
2. Background 241
2.1. Reasons for ensemble 241
2.2. Diversity and ensemble 241
2.3. Relationship between measures of diversity and combination method 243
2.4. Measures of diversity 243
2.5. Microarray data 244
3. Ensemble Classification System (ECS) Design 245
3.1. ECS overview 245
3.2. Feature subset selection 247
3.3. Base classifiers 248
3.4. Combination strategy 249
4. Experiments 250
4.1. Experimental datasets 250
4.2. Experimental results 252
5. Conclusion and Further Work 254
References 255

Chapter 12 An Automated Method for Cell Phase Identification in High Throughput Time-Lapse Screens 257
1. Introduction 258
2. Nuclei Segmentation and Tracking 259
3. Cell Phase Identification 260
3.1. Feature calculation 260
3.2. Identifying cell phase 262
3.3. Correcting cell phase identification errors 265
4. Experimental Results 266
5. Conclusion 272
References 272

Chapter 13 Inference of Transcriptional Regulatory Networks Based on Cancer Microarray Data 275
1. Introduction 275
2. Subnetworks and Transcriptional Regulatory Networks Inference 277
2.1. Inferring subnetworks using z-score 277
2.2. Inferring subnetworks based on graph theory 278
2.3. Inferring subnetworks based on Bayesian networks 279
2.4. Inferring transcriptional regulatory networks based on integrated expression and sequence data 283
3. Multinomial Probit Regression with Baysian Gene Selection 284
3.1. Problem formulation 284
3.2. Bayesian variable selection 286
3.3. Bayesian estimation using the strongest genes 288
3.4. Experimental results 289
4. Network Construction Based on Clustering and Predictor Design 293
4.1. Predictor construction using reversible jump MCMC annealing 293
4.2. CoD for predictors 295
4.3. Experimental results on a Myeloid line 296
5. Concluding Remarks 298
References 299

Chapter 14 Data Mining in Biomedicine 305
1. Introduction 305
2. Predictive Model Construction 306
2.1. Derivation of unsupervised models 307
2.2. Derivation of supervised models 311
3. Validation 316
4. Impact Analysis 318
5. Summary 319
References 319