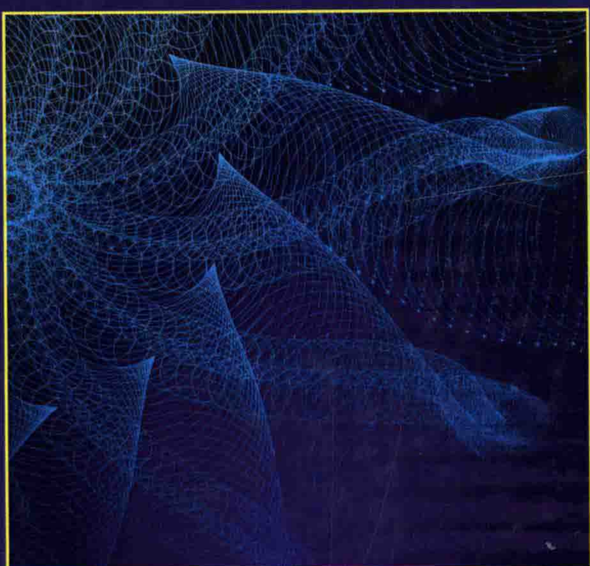
Wiley Series in Bioinformatics: Computational Techniques and Engineering

Yi P[illegible]n and Albert Y. Zomaya, Series Editors

Classification Analysis of DNA Microarrays



LEIF E. PETERSON

IEEE IEEE computer society WILEY

CLASSIFICATION ANALYSIS OF DNA MICROARRAYS

LEIF E. PETERSON
Director, Center for Biostatistics, The Methodist Hospital Research Institute, Houston, Texas
Associate Professor of Public Health, Weill Cornell Medical College, Cornell University, New York

WILEY | IEEE computer society

Cover design: John Wiley & Sons, Inc.
Cover illustration: © Carlos Olivares/iStockphoto

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Peterson, Leif E.
Classification analysis of DNA microarrays / Leif E. Peterson
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-17081-6 (cloth)

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CLASSIFICATION ANALYSIS OF DNA MICROARRAYS

Wiley Series in

Bioinformatics: Computational Techniques and Engineering

Bioinformatics and computational biology involve the comprehensive application of mathematics, statistics, science, and computer science to the understanding of living systems. Research and development in these areas require cooperation among specialists from the fields of biology, computer science, mathematics, statistics, physics, and related sciences. The objective of this book series is to provide timely treatments of the different aspects of bioinformatics spanning theory, new and established techniques, technologies and tools, and application domains. This series emphasizes algorithmic, mathematical, statistical, and computational methods that are central in bioinformatics and computational biology.

Series Editors: **Professor Yi Pan** pan@cs.gsu.edu **Professor Albert Y. Zomaya** zomaya@it.usyd.edu.au

---

Knowledge Discovery in Bioinformatics: Techniques, Methods and Applications / Xiaohua Hu & Yi Pan

Grid Computing for Bioinformatics and Computational Biology / Albert Zomaya & El-Ghazali Talbi

Analysis of Biological Networks / Björn H. Junker & Falk Schreiber

Bioinformatics Algorithms: Techniques and Applications / Ion Mandoiu & Alexander Zelikovsky

Machine Learning in Bioinformatics / Yanqing Zhang & Jagath C. Rajapakse

Biomolecular Networks / Luonan Chen, Rui-Sheng Wang, & Xiang-Sun Zhang

Computational Systems Biology / Huma Lodhi

Computational Intelligence and Pattern Analysis in Biology Informatics / Ujjwal Maulik, Sanghamitra Bandyopadhyay, & Jason T. Wang

Mathematics of Bioinformatics: Theory, Practice, and Applications / Matthew He & Sergey Petoukhov

Introduction to Protein Structure Prediction: Methods and Algorithms / Huzefa Rangwala & George Karypis

Mathematical and Computational Methods in Biomechanics of Human Skeletal Systems: An Introduction Jiri Nedoma & Jiri Stehlik

Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging / Pradipta Maji & Sankar K. Pal

Data Management of Protein Interaction Networks / Mario Cannataro & Pietro Hiram Guzzi

Classification Analysis of DNA Microarrays / Leif E. Peterson

To Susan and Evan

PREFACE

Classification analysis is a long-established technique that originated from the areas of statistics and pattern recognition, and is a broad rubric that subsumes *class discovery* and *class prediction*. Earlier forms of class discovery methods of statistical origin include natural grouping techniques such as hierarchical cluster analysis, K-means cluster analysis, and Gaussian mixture models employing the expectation–maximization algorithm. Conversely, pattern recognition-based class discovery techniques include self-organizing maps (Kohonen networks), neural gas, and fuzzy K-means cluster analysis. Class prediction methods that originated in statistics include linear discriminant analysis and logistic regression, while prototype learning, artificial neural networks, and swarm intelligence are more popular in pattern recognition. The main difference between the statistical and pattern recognition approaches is clear; the statistical methods tend to depend more on large sample Gaussian inference, while pattern recognition approaches tend to depend more on distribution-free heuristics employed in machine learning, evolutionary algorithms, or computational intelligence methods.

This book introduces the reader to a variety of statistical, machine learning, and computational intelligence classification algorithms that have been in use for several decades, as well as more recently developed algorithms based on fuzzy methods (soft computing), evolutionary algorithms, and swarm intelligence. Dimensional reduction and text mining for concept and document clustering are also covered to introduce the reader to information retrieval.

The layout of the book is divided into three parts. Part I, on class discovery, includes Chapters 1–9, which address various techniques for identifying the cluster structure of a dataset. Part II, on dimensional reduction, includes Chapters 10 and 11, which focus on linear and nonlinear approaches for reducing the dimensions of a dataset. Part III, on class prediction, covers Chapters 12–28, which present numerous approaches for predicting class membership of test objects after algorithm training is performed. There are also five appendixes, which cover probability, matrix algebra, mathematical functions, statistical primitives, and probability distributions. These are followed by a glossary of the symbols and notation presented throughout the book.

Chapter 1 summarizes class discovery, novel diagnostic classes, comorbidity and disease overlap, outliers and heterogeneity, class prediction, rules of thumb, and descriptions of the nine microarray datasets used throughout this book. Chapter 2 introduces a crisp *K*-means cluster analysis algorithm, distance metrics, cluster validity to determine the optimal number of clusters, and cluster initialization. Chapter 3 describes use of fuzzification to develop membership functions, which enable the presentation of cluster weights for each microarray. Chapter 4 covers unsupervised cluster analysis using Kohonen networks [self-organizing maps (SOMs)] and the numerous uses of SOM for understanding cluster structure. The fundamental components of self-organizing maps such as neighborhood functions, best-matching units, component maps, and *U* matrices are also discussed. Chapter 5 introduces prototype learning and the exploration of the cluster structure of data through neural adaptive learning with prototypes. Chapter 6 covers agglomerative clustering, correlation and distance-based agglomeration, dendograms, and heatmaps. Chapter 7 addresses Gaussian mixture models and the expectation–maximization (EM) algorithm for clustering. Chapter 8 develops document and concept clustering via text mining. Methods discussed include stopping, stemming, hash tables, inverse document frequency, and concept vectors. Chapter 9 extends text mining with the use of *N* grams.

Chapter 10, which discusses linear dimensional reduction by eigenanalysis of the gene-by-gene or array-by-array correlation matrix, develops the concepts of principal component score coefficients, loadings, and principal component scores. Chapter 11 is presented as a means of distance metric learning and dimensional reduction from a nonlinear standpoint, and includes kernel principal component analysis (PCA), diffusion maps, Laplacian eigenmaps, local linear embedding, locality preserving projections, and Sammon mapping.

Chapter 12 presents various methods used for filtering genes in order to develop "optimal" gene lists. A review of variable types (continuous, nominal, ordinal, etc.) is provided, as well as several 2- and *k*-sample parametric and nonparametric statistical tests for identifying best-ranked genes. A sequential forward–reverse sequential floating method known as "greedy plus takeaway" (greedy PTA) is also introduced, which forms the basis for optimal gene sets used throughout the book. Chapter 13 reviews computational efficiency, confusion matrices and accuracy calculations, cross-validation, bootstrapping, ensemble classifier fusion, random oracles, sensitivity and specificity, receiver–operator characteristic (ROC) curves, and area under the curve (AUC). Chapter 14 presents a matrix algebra approach to multivariate regression in which dependent variables for

ABBREVIATIONS

AAO	all at once
ACO	ant colony optimization
AID	automatic interaction detection
ANN	artificial neural network(s)
APP	all possible pairs
AQE	average quantization error
ARD	automatic relevance detection
AUC	area under (the) curve
BI	business intelligence
BLOG	binary logistic regression
BMU	best matching unit
CART	classification and regression tress
CDF/PDF	cumulative/probability distribution (density) function
CCAAT	cytidine–cytidine–adenosine–adenosine–thymidine
CKM	crisp *K* means
CMF	covariance matrix filtering
CMSA	covariance matrix self-adaptation (CMSA-ES = CMSA evolution strategies)
CV	cross-validation (CV − 1 = leave-one-out cross-validation)
DM	diffusion map
DTC	decision tree classification
EM	expectation–maximization
EMV	ensemble majority voting (EMWV = ensemble weighted majority voting)
FDA	Fisher's discriminant analysis
FDR	false discovery rate
FKM	fuzzy *K* means
FP/FN	false positive/negative (FPR/FNR = false positive/negative rate; TP/TN = true positive/negative, TPR/TNR = true positive/negative rate)
FWER	familywise error rate
GA	genetic algorithm

GMM	Gaussian mixture model
GOE	Gaussian orthogonal ensemble
GOF	goodness of fit
HCA	hierarchical cluster analysis
HDF	Hierarchical data format
HSV	hue–saturation–value
ICA	independent component analysis
IRLS	iteratively reweighted least squares
KDE	kernel density estimation (KDPCA = kernel density PCA)
KNN	*K* nearest neighbor
KREG	kernel regression
LDA	linear discriminant analysis
LEM	Laplacian eigenmap(s)
LL	loglikelihood
LLE	local linear embedding
LOG	logistic regression
LOOCV	leave-one-out cross-validation
LPP	locality preserving projection(s)
LREG	linear regression
LVQ	learning vector quantization
MCMC	Markov chain Monte Carlo
MLP	multilayer perceptron
MOE	mixture of experts
MSE/MST	mean-square error/total
MSPC	mutative strategy parameter control
NBC	naïve Bayes classification
NG	neural gas (SNG/UNG = supervised/unsupervised NG)
NLML	nonlinear manifold learning
OOA	one against all
OOB	out of (the) bag
ORC	outlier removal clustering
PC	principal component [PCA = principal component(s) analysis]
PDLO	principal direction linear oracle
PSO	particle swarm optimization
PTA	(greedy) plus takeaway
PV	predictive value
QDA	quadratic discriminant analysis
RBF	radial basis function
RF	random forest(s)
RFE	recursive feature elimination
RGB	red-green-blue
RMT	random matrix theory
ROC	receiver–operator characteristic
ROI	return on investment

CONTENTS