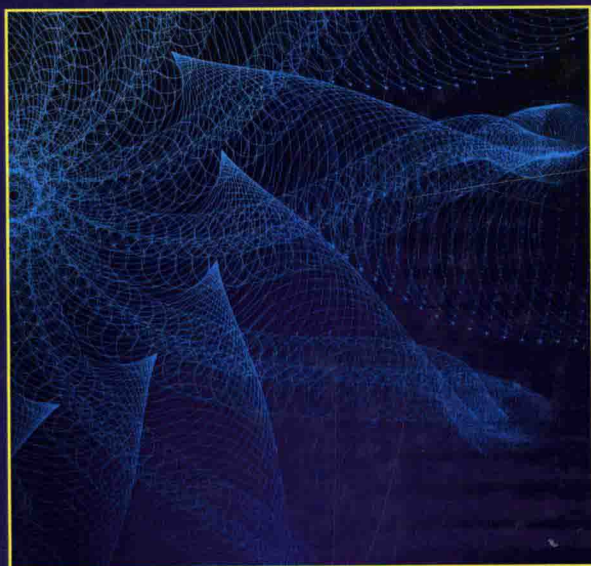


**Wiley Series in Bioinformatics:  
Computational Techniques and Engineering**

*John J. Schaeffer and Albert Y. Zomaya, Series Editors*

# Classification Analysis of DNA Microarrays



**LEIF E. PETERSON**



**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

Copyright © 2013 by John Wiley & Sons, Inc. All rights reserved

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

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at [www.copyright.com](http://www.copyright.com). Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

**Limit of Liability/Disclaimer of Warranty:** While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at [www.wiley.com](http://www.wiley.com).

***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**    **Professor Albert Y. Zomaya**  
pan@cs.gsu.edu    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, dendrograms, 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

Preface	xix
Abbreviations	xxiii
<b>1 Introduction</b>	<b>1</b>
1.1 Class Discovery	2
1.2 Dimensional Reduction	4
1.3 Class Prediction	4
1.4 Classification Rules of Thumb	5
1.5 DNA Microarray Datasets Used	9
References	11
 <b>PART I CLASS DISCOVERY</b>	 <b>13</b>
<b>2 Crisp K-Means Cluster Analysis</b>	<b>15</b>
2.1 Introduction	15
2.2 Algorithm	16
2.3 Implementation	18
2.4 Distance Metrics	20
2.5 Cluster Validity	24
2.5.1 Davies–Bouldin Index	25
2.5.2 Dunn’s Index	25
2.5.3 Intracuster Distance	26
2.5.4 Intercluster Distance	27
2.5.5 Silhouette Index	30
2.5.6 Hubert’s $\Gamma$ Statistic	31
2.5.7 Randomization Tests for Optimal Value of $K$	31
2.6 V-Fold Cross-Validation	35
2.7 Cluster Initialization	37
	<b>vii</b>

2.7.1	<i>K</i> Randomly Selected Microarrays	37
2.7.2	<i>K</i> Random Partitions	40
2.7.3	Prototype Splitting	41
2.8	Cluster Outliers	44
2.9	Summary	44
	References	45
<b>3</b>	<b>Fuzzy <i>K</i>-Means Cluster Analysis</b>	<b>47</b>
3.1	Introduction	47
3.2	Fuzzy <i>K</i> -Means Algorithm	47
3.3	Implementation	49
3.4	Summary	54
	References	54
<b>4</b>	<b>Self-Organizing Maps</b>	<b>57</b>
4.1	Introduction	57
4.2	Algorithm	57
4.2.1	Feature Transformation and Reference Vector Initialization	59
4.2.2	Learning	60
4.2.3	Conscience	61
4.3	Implementation	63
4.3.1	Feature Transformation and Reference Vector Initialization	63
4.3.2	Reference Vector Weight Learning	66
4.4	Cluster Visualization	67
4.4.1	Crisp <i>K</i> -Means Cluster Analysis	67
4.4.2	Adjacency Matrix Method	68
4.4.3	Cluster Connectivity Method	69
4.4.4	Hue–Saturation–Value (HSV) Color Normalization	69
4.5	Unified Distance Matrix ( <i>U</i> Matrix)	71
4.6	Component Map	71
4.7	Map Quality	73
4.8	Nonlinear Dimension Reduction	75
	References	79
<b>5</b>	<b>Unsupervised Neural Gas</b>	<b>81</b>
5.1	Introduction	81
5.2	Algorithm	82
5.3	Implementation	82

5.3.1	Feature Transformation and Prototype Initialization	82
5.3.2	Prototype Learning	83
5.4	Nonlinear Dimension Reduction	85
5.5	Summary	87
	References	88
<b>6</b>	<b>Hierarchical Cluster Analysis</b>	<b>91</b>
6.1	Introduction	91
6.2	Methods	91
6.2.1	General Programming Methods	91
6.2.2	Step 1: Cluster-Analyzing Arrays as Objects with Genes as Attributes	92
6.2.3	Step 2: Cluster-Analyzing Genes as Objects with Arrays as Attributes	94
6.3	Algorithm	96
6.4	Implementation	96
6.4.1	Heatmap Color Control	96
6.4.2	User Choices for Clustering Arrays and Genes	97
6.4.3	Distance Matrices and Agglomeration Sequences	98
6.4.4	Drawing Dendograms and Heatmaps	104
	References	105
<b>7</b>	<b>Model-Based Clustering</b>	<b>107</b>
7.1	Introduction	107
7.2	Algorithm	110
7.3	Implementation	111
7.4	Summary	116
	References	117
<b>8</b>	<b>Text Mining: Document Clustering</b>	<b>119</b>
8.1	Introduction	119
8.2	Duo-Mining	119
8.3	Streams and Documents	120
8.4	Lexical Analysis	120
8.4.1	Automatic Indexing	120
8.4.2	Removing Stopwords	121
8.5	Stemming	121
8.6	Term Weighting	121
8.7	Concept Vectors	124

8.8	Main Terms Representing Concept Vectors	124
8.9	Algorithm	125
8.10	Preprocessing	127
8.11	Summary	137
	References	137
<b>9</b>	<b>Text Mining: N-Gram Analysis</b>	<b>139</b>
9.1	Introduction	139
9.2	Algorithm	140
9.3	Implementation	141
9.4	Summary	154
	References	156
<b>PART II</b>	<b>DIMENSION REDUCTION</b>	<b>159</b>
<b>10</b>	<b>Principal Components Analysis</b>	<b>161</b>
10.1	Introduction	161
10.2	Multivariate Statistical Theory	161
10.2.1	Matrix Definitions	162
10.2.2	Principal Component Solution of $\mathbf{R}$	163
10.2.3	Extraction of Principal Components	164
10.2.4	Varimax Orthogonal Rotation of Components	166
10.2.5	Principal Component Score Coefficients	168
10.2.6	Principal Component Scores	169
10.3	Algorithm	170
10.4	When to Use Loadings and PC Scores	170
10.5	Implementation	171
10.5.1	Correlation Matrix $\mathbf{R}$	171
10.5.2	Eigenanalysis of Correlation Matrix $\mathbf{R}$	172
10.5.3	Determination of Loadings and Varimax Rotation	174
10.5.4	Calculating Principal Component (PC) Scores	176
10.6	Rules of Thumb For PCA	182
10.7	Summary	186
	References	187
<b>11</b>	<b>Nonlinear Manifold Learning</b>	<b>189</b>
11.1	Introduction	189
11.2	Correlation-Based PCA	190
11.3	Kernel PCA	191
11.4	Diffusion Maps	192

11.5	Laplacian Eigenmaps	192
11.6	Local Linear Embedding	193
11.7	Locality Preserving Projections	194
11.8	Sammon Mapping	195
11.9	NLML Prior to Classification Analysis	195
11.10	Classification Results	197
11.11	Summary	200
	References	203

## **PART III CLASS PREDICTION** **205**

### **12 Feature Selection** **207**

12.1	Introduction	207
12.2	Filtering versus Wrapping	208
12.3	Data	209
	12.3.1 Numbers	209
	12.3.2 Responses	209
	12.3.3 Measurement Scales	210
	12.3.4 Variables	211
12.4	Data Arrangement	211
12.5	Filtering	213
	12.5.1 Continuous Features	213
	12.5.2 Best Rank Filters	219
	12.5.3 Randomization Tests	236
	12.5.4 Multitesting Problem	237
	12.5.5 Filtering Qualitative Features	242
	12.5.6 Multiclass Gini Diversity Index	246
	12.5.7 Class Comparison Techniques	247
	12.5.8 Generation of Nonredundant Gene List	250
12.6	Selection Methods	254
	12.6.1 Greedy Plus Takeaway (Greedy PTA)	254
	12.6.2 Best Ranked Genes	258
12.7	Multicollinearity	259
12.8	Summary	270
	References	270

### **13 Classifier Performance** **273**

13.1	Introduction	273
13.2	Input–Output, Speed, and Efficiency	273
13.3	Training, Testing, and Validation	277

13.4	Ensemble Classifier Fusion	280
13.5	Sensitivity and Specificity	283
13.6	Bias	284
13.7	Variance	285
13.8	Receiver–Operator Characteristic (ROC) Curves	286
	References	295
<b>14</b>	<b>Linear Regression</b>	<b>297</b>
14.1	Introduction	297
14.2	Algorithm	299
14.3	Implementation	299
14.4	Cross-Validation Results	300
14.5	Bootstrap Bias	303
14.6	Multiclass ROC Curves	306
14.7	Decision Boundaries	308
14.8	Summary	310
	References	310
<b>15</b>	<b>Decision Tree Classification</b>	<b>311</b>
15.1	Introduction	311
15.2	Features Used	314
15.3	Terminal Nodes and Stopping Criteria	315
15.4	Algorithm	315
15.5	Implementation	315
15.6	Cross-Validation Results	318
15.7	Decision Boundaries	326
15.8	Summary	327
	References	329
<b>16</b>	<b>Random Forests</b>	<b>331</b>
16.1	Introduction	331
16.2	Algorithm	333
16.3	Importance Scores	334
16.4	Strength and Correlation	338
16.5	Proximity and Supervised Clustering	342
16.6	Unsupervised Clustering	345
16.7	Class Outlier Detection	348
16.8	Implementation	350
16.9	Parameter Effects	350
16.10	Summary	357
	References	358



<b>17</b>	<b>K Nearest Neighbor</b>	<b>361</b>
17.1	Introduction	361
17.2	Algorithm	362
17.3	Implementation	363
17.4	Cross-Validation Results	364
17.5	Bootstrap Bias	369
17.6	Multiclass ROC Curves	373
17.7	Decision Boundaries	374
17.8	Summary	377
	References	378
<b>18</b>	<b>Naïve Bayes Classifier</b>	<b>379</b>
18.1	Introduction	379
18.2	Algorithm	380
18.3	Cross-Validation Results	380
18.4	Bootstrap Bias	384
18.5	Multiclass ROC Curves	386
18.6	Decision Boundaries	386
18.7	Summary	389
	References	391
<b>19</b>	<b>Linear Discriminant Analysis</b>	<b>393</b>
19.1	Introduction	393
19.2	Multivariate Matrix Definitions	394
19.3	Linear Discriminant Analysis	396
19.3.1	Algorithm	397
19.3.2	Cross-Validation Results	397
19.3.3	Bootstrap Bias	401
19.3.4	Multiclass ROC Curves	402
19.3.5	Decision Boundaries	403
19.4	Quadratic Discriminant Analysis	403
19.5	Fisher's Discriminant Analysis	406
19.6	Summary	411
	References	412
<b>20</b>	<b>Learning Vector Quantization</b>	<b>415</b>
20.1	Introduction	415
20.2	Cross-Validation Results	417
20.3	Bootstrap Bias	417
20.4	Multiclass ROC Curves	426