

Contents

1	Introduction.....	1
1.1	Context.....	1
1.2	New Scientific Techniques, New Data Challenges in Life Sciences.....	8
1.3	The Ethics of Data Mining	10
1.4	Data Mining: Problems and Issues	12
1.5	From Data to Information: The Process.....	17
1.5.1	CRISP-DM.....	19
1.6	What Can Be Mined?.....	24
1.7	Standards and Terminologies.....	25
1.8	Interestingness	26
1.9	How Good Is Good Enough?.....	27
1.10	The Datasets Used in This Book	28
1.11	The Rest of the Story	28
1.12	What Have We Missed in the Introduction?	31
	References.....	31
2	Fundamental Concepts.....	33
2.1	Introduction	33
2.2	Data Mining Activities.....	37
2.3	Models	38
2.4	Input Representation.....	39
2.5	Missing Data.....	40
2.6	Does Data Expire?.....	41
2.7	Frequent Patterns.....	42
2.8	Bias.....	43
2.9	Generalization.....	45
2.10	Data Characterization and Discrimination.....	45
2.11	Association Analysis.....	46
2.12	Classification and Prediction.....	46
2.12.1	Relevance Analysis.....	47

2.13	Cluster Analysis.....	47
2.14	Outlier Analysis.....	48
2.15	Normalization.....	49
2.16	Dimensionality.....	49
2.17	Model Overfitting.....	51
2.18	Concept Hierarchies	52
2.19	Bias-Variance Decomposition.....	55
2.20	Advanced Techniques.....	55
2.20.1	Principal Component Analysis	56
2.20.2	Monte Carlo Strategies	56
2.21	Some Introductory Methods.....	60
2.21.1	The 1R Method	60
2.21.2	Decision Trees	63
2.21.3	Classification Rules.....	66
2.21.4	Association Rules.....	69
2.21.5	Relational Rules	72
2.21.6	Clustering	73
2.21.7	Simple Statistical Measures	74
2.21.8	Linear Models	74
2.21.9	Neighbors and Distance	76
2.22	Feature Selection.....	79
2.22.1	Global Feature Selection	80
2.22.2	Local Feature Selection.....	81
2.23	Conclusion	82
	References.....	82
3	Data Architecture and Data Modeling.....	85
3.1	Introduction.....	85
3.2	Data Modeling, a Whirlwind Tour.....	86
3.3	Qualitative and Quantitative Data.....	90
3.4	Interpretive Data.....	93
3.5	Sequence Data	93
3.6	Gene Ontology.....	95
3.7	Image Data.....	96
3.8	Text Data.....	97
3.9	Protein and Compound 3D Structural Data	97
3.10	Data Integration.....	102
3.10.1	Same Entity, Different Identifiers	102
3.10.2	Data in Different Formats	103
3.10.3	Default Values	103
3.10.4	Multiscale Data Integration.....	104
3.11	The Test Subject Area.....	106
3.12	The Transactional Data Model, Operational Data Store, Data Warehouse, and Data Mart.....	110

3.13	Modeling Data with a Time Dimension.....	112
3.14	Historical Data and Archiving: Another Temporal Issue?.....	113
3.15	Data Management.....	114
3.16	Agile Modeling, Adaptive Data Warehouses, and Hybrid Models.....	115
3.17	Gene Ontology Database	116
3.18	Broadening Our Concept of Data: Images, Structures, and so on.....	119
3.19	Conclusion	122
	References.....	122
4	Representing Data Mining Results.....	125
4.1	Introduction.....	125
4.2	Tabular Representations	126
4.3	Decision Tables, Decision Trees, and Classification Rules	127
4.4	Classification and Regression Trees (CARTs).....	129
4.4.1	Process	133
4.4.2	Cross-validation.....	143
4.4.3	Summary	144
4.5	Association Rules.....	145
4.5.1	Context.....	145
4.5.2	The Apriori Algorithm.....	148
4.5.3	Rules, Rules, and More Rules	150
4.6	Some Basic Graphical Representations.....	152
4.6.1	Representing the Instances Themselves	152
4.6.2	Histograms	153
4.6.3	Frequency Polygrams	154
4.6.4	Data Plots	154
4.7	Representing the Output from Statistical Models.....	158
4.7.1	Boxplots.....	159
4.7.2	Scatterplots.....	161
4.7.3	Q-Q Plots.....	162
4.8	Heat Maps.....	166
4.9	Position Maps.....	170
4.9.1	Microbial Genome Viewer	170
4.9.2	GenomePlot	171
4.9.3	GenoMap.....	171
4.9.4	Circular Genome Viewer	171
4.10	Genotype Visualization.....	171
4.11	Network Visualizations.....	172
4.12	Some Tools.....	175
4.12.1	Cytoscape	176
4.12.2	TreeView: Phylogenetic Tree Visualization	177

4.12.3	Systems Biology Markup Language	178
4.12.4	R and RStudio	181
4.12.5	Weka.....	183
4.12.6	Systems Biology Graphical Notation.....	183
4.13	Conclusion	185
	References.....	189
5	The Input Side of the Equation.....	191
5.1	Data Preparation	194
5.2	Internal Consistency	195
5.3	External Consistency.....	197
5.4	Standardization.....	198
5.5	Transformation.....	198
5.5.1	Transforming Categorical Attributes	199
5.6	Normalization	200
5.6.1	Averaging	200
5.6.2	Min-Max Normalization	200
5.7	Denormalization	201
5.7.1	Prejoining Tables	201
5.7.2	Report Tables	202
5.7.3	Mirroring Tables	202
5.7.4	Splitting Tables	203
5.7.5	Combining Tables	203
5.7.6	Copying Data Across Entities	203
5.7.7	Repeating Groups	204
5.7.8	Derived Data	205
5.7.9	Hierarchies of Data	205
5.8	Extract, Transform, and Load (ETL).....	206
5.8.1	Integrating External Data	207
5.8.2	Gene/Protein Identifier Cross-Referencing	208
5.9	Aggregation and Calculation	209
5.9.1	Aggregation	209
5.9.2	Calculation	211
5.10	Noisy Data	211
5.10.1	Regression	212
5.10.2	Binning.....	212
5.10.3	Clustering	215
5.11	Metadata	215
5.11.1	Attribute Type	215
5.11.2	Attribute Role.....	216
5.11.3	Attribute Description	217
5.12	Missing Values.....	217
5.12.1	Other Issues with Missing Data	223
5.13	Data Reduction.....	224

5.13.1	Attribute Subsets.....	224
5.13.2	Data Aggregation	225
5.13.3	Dimensional Reduction.....	225
5.13.4	Alternate Representations	225
5.13.5	Generalization	226
5.13.6	Discretization	227
5.14	Refreshing the Dataset.....	228
5.14.1	Adding Data to Our Dataset.....	229
5.14.2	Removing Data from Our Dataset.....	229
5.14.3	Correcting the Data in Our Dataset	230
5.14.4	Data Changes that Significantly Affect Our Dataset.....	232
5.15	Conclusion	233
	References.....	233
6	Statistical Methods.....	235
6.1	Introduction	235
6.1.1	Basic Statistical Methods	237
6.2	Statistical Inference and Hypothesis Testing.....	239
6.2.1	What Constitutes a “Good” Hypothesis?	241
6.2.2	The Null Hypothesis	242
6.2.3	p Value	243
6.2.4	Type I and Type II Errors.....	245
6.2.5	An Example: Comparing Two Groups, the t Test.....	246
6.2.6	Back to Square One?.....	252
6.2.7	Some Other Hypothesis Testing Methods	254
6.2.8	Multiple Comparisons and Multiple-Testing Correlation.....	259
6.3	Measures of Central Tendency, Variability, and Other Descriptive Statistics	266
6.3.1	Location	266
6.3.2	Variability	268
6.3.3	Heterogeneity	269
6.3.4	Concentration	270
6.3.5	Asymmetry	272
6.3.6	Kurtosis	273
6.3.7	Same Means, Different Variances – Different Variances, Same Means	274
6.4	Frequency Distributions	276
6.5	Confidence Intervals	278
6.5.1	Continuous Data Confidence Levels.....	278
6.6	Regression Analysis	282
6.6.1	Linear Regression	283
6.6.2	Correlation Coefficient.....	289
6.6.3	Multiple Linear Regression	289

6.7	Maximum Likelihood Estimation Method	290
6.7.1	Illustration of the MLE Method	291
6.7.2	Use in Data Mining	292
6.8	Maximum A Posteriori (MAP) Estimation.....	293
6.9	Enrichment Analysis.....	294
6.10	False Discover Rate (FDR).....	297
6.11	Statistical Significance and Clinical Relevance	300
6.12	Conclusion	300
	References.....	301
7	Bayesian Statistics	303
7.1	Introduction	303
7.2	Bayesian Formulations.....	304
7.2.1	Bayes' Theorem.....	309
7.3	Assigning Probabilities, Odds Ratios, and Bayes Factor.....	312
7.3.1	Probability Assignment	312
7.3.2	Odds Ratios	314
7.3.3	Bayes Factor	314
7.4	Putting It All Together.....	315
7.5	Bayesian Reasoning.....	322
7.5.1	A Simple Illustration.....	324
7.6	Bayesian Classification	327
7.7	Bayesian Belief Networks	337
7.8	Parameter Estimation Methods.....	340
7.9	Multiple Datasets	342
7.10	Hidden Markov Models (HMM)	342
7.11	Conditional Random Field (CRF).....	351
7.12	Array Comparative Genomic Hybridization	355
7.13	Conclusion	360
	References.....	360
8	Machine-Learning Techniques	363
8.1	Introduction	363
8.1.1	Missing and Erroneous Data	366
8.2	Measure of Similarity.....	368
8.3	Supervised Learning	370
8.3.1	Classification Learning.....	371
8.4	Unsupervised Learning	372
8.4.1	Association Learning.....	372
8.4.2	Clustering	372
8.5	Semisupervised Learning	373
8.5.1	Expectation Maximization	374
8.5.2	Cotraining	377
8.5.3	Graph-Based SSL Methods	381
8.5.4	Is This Type of Approach Always Helpful?.....	383

8.6	Kernel Learning Methods.....	383
8.7	Let's Break for Some Examples.....	385
8.7.1	String and Tree Matching.....	386
8.7.2	Protein Structure Classification and Prediction	386
8.8	Support Vector Machines.....	389
8.8.1	Gene Expression Analysis	394
8.9	Artificial Neural Networks.....	399
8.9.1	Introduction	399
8.9.2	Neurons	399
8.9.3	Neuronal Functions	401
8.9.4	Encoding the Input.....	406
8.9.5	Training Methods	406
8.9.6	ANN Architectures	424
8.9.7	Application to Data Mining.....	424
8.10	Reinforcement Learning	425
8.11	Some Other Techniques	425
8.11.1	Random Walk, Diffusion Map, and Spectral Clustering.....	426
8.11.2	Network (Graph)-Based Analysis	431
8.11.3	Network Motif Discovery	431
8.11.4	Binary Tree Algorithm in Drug Target Discovery Studies.....	436
8.11.5	Petri Nets	437
8.11.6	Boolean and Fuzzy Logic	440
8.12	Conclusion	448
	References.....	449
9	Classification and Prediction.....	455
9.1	Introduction	455
9.2	Data Preparation	459
9.3	Linear Regression.....	460
9.4	Decision Trees	463
9.5	1R	466
9.6	Nearest Neighbor	469
9.7	Bayesian Modeling	471
9.8	Neural Networks.....	477
9.9	k-Means.....	481
9.10	Distance Measures.....	486
9.11	Measuring Accuracy.....	489
9.11.1	Classifiers	489
9.11.2	Predictors	492
9.11.3	Evaluating Accuracy.....	493
9.11.4	Improving Classifier/Predictor Accuracy.....	496
9.12	Conclusion	497
	References.....	498

10 Informatics	501
10.1 Introduction	501
10.1.1 Sources of Genomic and Proteomic Data	503
10.2 Data Integration	504
10.2.1 Integrating Annotations for a Common Sequence	505
10.2.2 Data in Different Formats	505
10.3 Tools and Databases	508
10.3.1 Programming Languages and Environments	511
10.4 Standardization	512
10.5 Microarrays	513
10.5.1 Challenges and Opportunities	514
10.5.2 Classification	515
10.5.3 Gene Selection	515
10.6 Finding Motifs	516
10.6.1 Regular Expressions	516
10.7 Analyzing DNA	521
10.7.1 Pairwise Sequence Alignment	521
10.7.2 Multiple Alignment	527
10.7.3 Trees	531
10.8 Conclusion	538
References	541
11 Systems Biology	543
11.1 What Is This Systems Biology of Which You Speak?	543
11.2 Biological Networks	547
11.3 How Much Biology Do We Need?	553
11.3.1 Biological Processes As Ordered Sequences of Events	553
11.4 But, We Do Need Some Graph Theory	554
11.4.1 Data Mining by Navigation	561
11.4.2 So...How Are We Going to Cram All This into a Single Chapter?	562
11.5 Gene Ontology and Microarray Databases	563
11.5.1 Gene Ontology (GO)	564
11.5.2 Microarray Databases	565
11.6 Text Mining	565
11.7 Some Core Problems and Techniques	571
11.7.1 Shotgun Fragment Assembly	572
11.7.2 The BioCreAtIvE Initiative	575
11.8 Data Mining in Systems Biology	576
11.8.1 Network Analysis	577
11.9 Novel Initiatives	577
11.9.1 Data Mining Promises to Dig Up New Drugs	577
11.9.2 Temporal Interactions Among Genes	578

11.10	The Cloud.....	579
11.11	Where to Next?.....	580
11.12	Have We Left Anything Out? Boy, Have We Ever	580
	References.....	581
12	Let's Call It a Day.....	585
12.1	We've Covered a Lot, But Not Really.....	585
12.2	When Two Models Are Better Than One.....	586
12.3	The Most Widely Used Algorithms?.....	587
12.4	Documenting Your Process	588
12.5	Where To From Here?.....	589
	References.....	590
	Appendix A.....	593
	Appendix B.....	603
	Appendix C.....	623
	Appendix D.....	627
	References.....	629
	Index.....	631



<http://www.springer.com/978-1-58829-942-0>

Introduction to Data Mining for the Life Sciences

Sullivan, R.

2012, XVIII, 638 p., Hardcover

ISBN: 978-1-58829-942-0

A product of Humana Press