



Data Mining: Opportunities and Challenges

Table of Contents

Preface	vii
<i>John Wang, Montclair State University, USA</i>	
Chapter I	
A Survey of Bayesian Data Mining	1
<i>Stefan Arnborg, Royal Institute of Technology and Swedish Institute of Computer Science, Sweden</i>	
Chapter II	
Control of Inductive Bias in Supervised Learning Using Evolutionary Computation: A Wrapper-Based Approach	27
<i>William H. Hsu, Kansas State University, USA</i>	
Chapter III	
Cooperative Learning and Virtual Reality-Based Visualization for Data Mining	55
<i>Herna Viktor, University of Ottawa, Canada</i>	
<i>Eric Paquet, National Research Council, Canada</i>	
<i>Gys le Roux, University of Pretoria, South Africa</i>	
Chapter IV	
Feature Selection in Data Mining	80
<i>YongSeog Kim, University of Iowa, USA</i>	
<i>W. Nick Street, University of Iowa, USA</i>	
<i>Filippo Menczer, University of Iowa, USA</i>	
Chapter V	
Parallel and Distributed Data Mining through Parallel Skeletons and Distributed Objects	106
<i>Massimo Coppola, University of Pisa, Italy</i>	
<i>Marco Vanneschi, University of Pisa, Italy</i>	
Chapter VI	
Data Mining Based on Rough Sets	142
<i>Jerzy W. Grzymala-Busse, University of Kansas, USA</i>	
<i>Wojciech Ziarko, University of Regina, Canada</i>	

Chapter VII	
The Impact of Missing Data on Data Mining	174
<i>Marvin L. Brown, Hawaii Pacific University, USA</i>	
<i>John F. Kros, East Carolina University, USA</i>	
Chapter VIII	
Mining Text Documents for Thematic Hierarchies Using Self-Organizing Maps .	199
<i>Hsin-Chang Yang, Chang Jung University, Taiwan</i>	
<i>Chung-Hong Lee, Chang Jung University, Taiwan</i>	
Chapter IX	
The Pitfalls of Knowledge Discovery in Databases and Data Mining	220
<i>John Wang, Montclair State University, USA</i>	
<i>Alan Oppenheim, Montclair State University, USA</i>	
Chapter X	
Maximum Performance Efficiency Approaches for Estimating Best Practice Costs	239
<i>Marvin D. Troutt, Kent State University, USA</i>	
<i>Donald W. Gribbin, Southern Illinois University at Carbondale, USA</i>	
<i>Murali S. Shanker, Kent State University, USA</i>	
<i>Aimao Zhang, Georgia Southern University, USA</i>	
Chapter XI	
Bayesian Data Mining and Knowledge Discovery	260
<i>Eitel J. M. Lauria, State University of New York, Albany, USA, Universidad del Salvador, Argentina</i>	
<i>Giri Kumar Tayi, State University of New York, Albany, USA</i>	
Chapter XII	
Mining Free Text for Structure	278
<i>Vladimir A. Kulyukin, Utah State University, USA</i>	
<i>Robin Burke, DePaul University, USA</i>	
Chapter XIII	
Query-By-Structure Approach for the Web	301
<i>Michael Johnson, Madonna University, USA</i>	
<i>Farshad Fotouhi, Wayne State University, USA</i>	
<i>Sorin Draghici, Wayne State University, USA</i>	
Chapter XIV	
Financial Benchmarking Using Self-Organizing Maps—Studying the International Pulp and Paper Industry	323
<i>Tomas Eklund, Turku Centre for Computer Science, Finland</i>	
<i>Barbro Back, Åbo Akademi University, Finland</i>	
<i>Hannu Vanharanta, Pori School of Technology and Economics, Finland</i>	
<i>Ari Visa, Tampere University of Technology, Finland</i>	

Chapter XV	
Data Mining in Health Care Applications	350
<i>Fay Cobb Payton, North Carolina State University, USA</i>	
Chapter XVI	
Data Mining for Human Resource Information Systems	366
<i>Lori K. Long, Kent State University, USA</i>	
<i>Marvin D. Troutt, Kent State University, USA</i>	
Chapter XVII	
Data Mining in Information Technology and Banking Performance	382
<i>Yao Chen, University of Massachusetts at Lowell, USA</i>	
<i>Joe Zhu, Worcester Polytechnic Institute, USA</i>	
Chapter XVIII	
Social, Ethical and Legal Issues of Data Mining	395
<i>Jack S. Cook, Rochester Institute of Technology, USA</i>	
<i>Laura L. Cook, State University of New York at Geneseo, USA</i>	
Chapter XIX	
Data Mining in Designing an Agent-Based DSS	421
<i>Christian Böhm, GIDSATD-UTN-FRSF, Argentina</i>	
<i>Maria Rosa Galli, GIDSATD-UTN-FRSF and INGAR-CONICET, Argentina</i>	
<i>Omar Chiotti, GIDSATD-UTN-FRSF and INGAR-CONICET, Argentina</i>	
Chapter XX	
Critical and Future Trends in Data Mining: A Review of Key Data Mining Technologies/Applications	437
<i>Jeffrey Hsu, Fairleigh Dickinson University, USA</i>	
About the Authors	453
Index	462

Index

A

agent-based architecture 421
 agent-based DSS 421
 agents 440
 apriori 108
 artificial intelligence (AI) 447
 artificial neural networks (ANNs) 84
 association rules 108, 176, 188
 attribute subset selection 36
 attribute-based approaches 30
 automated problem decomposition 28
 automated relevance determination 33
 automatic category hierarchy generation
 207
 automatic category theme identification
 211
 automatic generation of categories 200
 automatic interaction detection 176
 autoregressive moving average (ARMA)
 41

B

bagging 96
 banking performance 382
 banks 383
 Bayes factor 3
 Bayes' theorem 260
 Bayesian analysis 6
 Bayesian approach 263
 Bayesian belief networks (BBNs) 260,
 269
 Bayesian classification 260
 Bayesian classifier 267

Bayesian methods 260
 Bayesian data mining 1
 behavioral factors 355
 Beowulf clusters 139
 Bernoulli distribution 264
 beta distribution 4, 264
 bioinformatics 445
 block of a variable-value pair 150
 blocks 148
 bookmark organizer (BO) 441
 boosting 96
 boundary region 156
 business intelligence 439

C

c-elementary sets 154
 C4.5 108
 candidate distribution 118
 CART® 176
 case substitution 182
 cases base 428
 categorical data 6
 categorization 441
 categorization agents 440
 category hierarchy 203
 category themes 200
 causal relationships 269
 causality 13
 CHAID 176
 chi-square automatic interaction
 detection 176
 child links 305
 Children's Online Privacy Protection Act
 (COPPA) 413

- Chinese corpus 201
- chordal graphs 15
- CLaP 308
- classification 375, 438
- classification table 155
- classification trees 176
- classifier fusion 30
- cliques 15
- clustering 33, 89, 203, 375, 438, 441
- co-citation analysis 440
- cold deck imputation 183
- collaborative virtual environments (CVEs) 73
- collective data-mining (CDM) 441
- community health information networks (CHINs) 350, 352
- competitive clustering 30
- complete case approach 181
- composite hypothesis 4
- composite learning 32
- computation grain 109
- computational grid resources 110
- concepts 148
- conditional entropy 38
- confounder 14
- conjugate prior 264
- constraint-based DM 444
- constructive induction 33
- convolutional code 38
- cooperative learning 55
- coordination language 111
- CORBA-operated software 110
- core point 128
- count distribution 118
- credible set 3
- crisis monitoring 28
- cross-validation 228
- curse of dimensionality 31
- customer (patient) profiling 352
- customer relationship management (CRM) 89
- customer targeting 84
- customized-usage tracking 439
- D**
- data analysis 367
- data cleaning 60
- data constraints 444
- data distribution 118
- data dredging 226
- data envelopment analysis (DEA) 383
- data hypercube 439
- data imputation 181
- data inconsistency 178
- data integration 61
- data matrix 6
- data mining (DM) 55, 107, 220, 260, 350, 351, 369, 382, 395
- data mining techniques 59, 367
- data mining tools 324
- data missing at random 175, 179
- data missing completely at random 175, 179
- data parallel 112
- data preprocessing 60
- data repository 59
- data selection 61
- data servers 110
- data to knowledge (D2K) 28
- data transformation 61
- data transport layers 110
- data visualization 235
- data warehouses 370
- database (DB) view 438
- database management support (DBMS) 107
- datamarts 370
- DBSCAN 108
- DEA efficient frontier 385
- decision matrices 159
- decision tree analysis 375
- decision tree inducers 28
- decision tree induction 108
- decision trees (DTs) 122, 188, 233
- decision-support system 421
- decomposition of input 32
- Dempster-Shafer theory 6
- density-based clustering 128
- design patterns 111
- digital libraries 442
- dimension/level constraints 445
- directed acyclic graph 269

Dirichlet distribution 10
 dirty data 177
 disaster planning 235
 discriminating analysis 424
 discriminating function weights 425
 distributed and collective DM 441
 distributed computation 106
 distributed data mining (DDM) 108, 441
 distributed hypertext resource discovery 442
 distributed memory architectures 108
 divide-and-conquer (D&C) 122
 DNA 446
 document cluster map (DCM) 201, 204
 document semantics 320
 domains dimension 428
 dynamic DSS 433

E

e-health models 351
 economic dimensions 354
 efficiency approaches 239
 efficient frontier 386
 electronic commerce 350
 Electronic Communications Privacy Act 411
 Electronic Funds Transfer Act 410
 elementary sets 154
 employee recruitment support 378
 employee training evaluation 379
 ensemble 81
 ensemble learning 33
 estimation criterion quality issues 249
 estimation maximization (EM) algorithm 273
 ethical issues 398
 evolutionary algorithms (EAs) 81, 84
 evolutionary computation 27
 evolutionary computation—genetic algorithms (GA) 34
 evolutionary computation genetic programming (GP) 34
 evolutionary local selection algorithms (ELSA) 81

expectation maximization (EM) algorithm 41
 expectation maximization (EM) method 17
 experimental design 5
 expert system for environmental protection (ESEP) 167
 external object library 108
 external objects (EO) 114
 extraction rules 438

F

factor analysis 30
 Fair Credit Reporting Act 408
 farm skeleton 127
 feature (attribute) partitioning 28
 feature extraction 441
 feature selection 81
 feature subset selection 27, 36
 feature vector 306
 feedforward layer 315
 financial benchmarking 323
 financial competitor benchmarking 324
 financial performance 383
 flat-file mining 109
 fraud detection 352
 frequent sets 116
 frequentist 263
 fuzzy logic (FL) 234

G

GeneCards 446
 general access-pattern tracking 439
 genetic algorithms (GAs) 177, 234
 genetic ensemble feature selection (GEFS) 97
 Gibbs sampler 275
 global covering 149
 global knowledge map 38
 global positioning systems (GPS) 31
 government policies 354
 Gramm-Leach-Bliley Act of 1999 412
 graphical model 7, 269

H

hash tree 118

- Health Insurance Portability and Accountability Act (HIPAA) 411
 - hierarchical mixture estimation 33
 - hierarchical mixture of experts (HME) 37
 - high-level optimization systems 27
 - high-level parallel programming 107
 - Hilbert space-filling curve 138
 - horizontal representation 115
 - hot deck imputation 176, 183
 - human brain 22
 - human resource (HR) 366
 - human resource information systems 368
 - human resources data 367
 - human-computer interaction 441
 - hyperlink and inter-document structures 442
 - hyperlinks 442
 - HyPursuit 440
- I**
- imputation methods 176
 - in-core techniques 107
 - inaccurate classifications 178
 - incomplete data 175
 - inconsistent data 147
 - indiscernibility relation 148
 - inductive bias 27, 29
 - inference 270
 - information access 440
 - information filtering 440
 - information gain (IG) 123
 - information quality 355
 - information retrieval 440
 - information retrieval system 279
 - information retrieval view 438
 - information sharing 355
 - information technology (IT) 382
 - information technology infrastructure (ITIS) 222
 - integration 63
 - intelligent search agents 440
 - interaction 62
 - interestingness constraints 445
 - intermediate concepts 30
 - international laws 414
 - inverse probability 3
- J**
- joint probability distribution 270
- K**
- k-itemset 116
 - k-means clustering 30
 - Karhunen-Loève transforms 30
 - knowledge chunks 55
 - knowledge discovery (KD) 28, 81, 176
 - knowledge discovery in databases (KDD) 220
 - knowledge-type constraints 444
- L**
- l-negative region 155
 - Laplace estimator 5, 269
 - large-scale KD problems 30
 - latent variables 18
 - lattice structure 116
 - learning from examples using rough sets (LERS) 145
 - learning process 421
 - legal issues 408
 - LERS classification system 146
 - load-balancing 439
 - local covering 151
 - lower approximation 151
 - lower limit l 155
- M**
- machine learning 279, 303
 - machine learning in Java (MLJ) 28
 - machine learning techniques 305
 - marginal cost-oriented basic cost models 254
 - market basket analysis 116
 - Markov Chain Monte Carlo methods (MCMC) 266
 - Markov chain simulation 274
 - Markov chains 274
 - matching factor 167
 - mathematical probability 2
 - Maximum A Posteriori (MAP) 5

maximum performance efficiency (MPE)
239, 244

mean 5

mean estimate 5

mean substitution 182

median 5

medical field 168

medical imaging 443

memory form 38

memory hierarchy 108

meta-evolutionary ensembles (MEE)
81, 97

metric-based model selection 32

metropolis algorithm 275

minimal complex 150

minimum confidence 177

minimum description length (MDL) 27

minimum support 116

mining free text for content markers
280

mixture model 32

mixture prior 10

model selection 28, 31, 37

model-based procedures 186

modular parity 36

moving average (MA) 41

MPI 111

multimedia DM 443

multimedia mining 442

multiple imputation 186

N

Naive Bayes 267

natural language 169

natural language processing 279

navigation 443

nearest neighbor methods 176

neighboring neurons 204

network transmission 439

neural network net query-by-structure
308

neural networks (NNs) 177, 232, 305,
324, 376

noise 231

non-ignorable missing data 180

non-monotonic reasoning 6

normal variation 4

normative claim 6

O

objective probability 5

online analytical mining architecture
(OLAM) 445

online catalogues 442

online information DBs 442

out-of-core techniques 107

outliers 176, 231

outliers treated as missing data 180

output layer 326

over-prediction 185

P

p-elementary sets 148

p-value 5

parallel computation 106

parallel data mining (PDM) 108

parallel programming 110

parameterized model 4

Pareto front 81

Pareto optimization 81

partial matching 167

partitioning 30

pattern recognition 177

patterns obtention 430

PDBSCAN 134

performance evaluation 384

periodicity analysis 444

permissible edge 15

personalization 439

personalized Web agents 440

phenomenal DM 445

pipe skeleton 112

posterior density 3

posterior odds 3

posterior probability 260

presentation consistency 280

presentation criteria 302

presentation similarity 280

principal components analysis 30

prior distribution 264

prior odds 3

privacy 397

probabilistic decision table 156
 probabilistic dependency 157
 probability 260
 problem definition 373
 profitability analysis 353
 programming systems product 29
 prototype vectors 314
 pulp and paper industry 326
 push or pull factors 355

Q

qualitative measures 327
 quality of lower approximation 152
 quality of upper approximation 153
 quantitative data 325
 query-by-structure approach 302, 304

R

R*-tree 129
 radial-basis functions 32
 random subspace method (RSM) 96
 random walk 274
 randomization 228
 real-time decision-making 28
 recommender systems 28
 recreational vehicle (RV) 84
 recurrent layer 316
 regression 377
 regression imputation 184
 rejection region 5
 rel-infons 306
 relative reduct 149
 relevance determination 31
 remote sensing 444
 representation 62
 representation bias 45
 resampling 228
 restricted parallel programming 110
 retention management 353
 return on investment (ROI) 222
 Right to Financial Privacy Act 411
 rough region 156
 rough set theory 147
 rule constraints 446
 rule of succession 10

S

sampling 227
 scalability 224, 265
 scalable data-mining solutions for free-text 281
 scatter plots 62
 schizophrenia 22
 search engines 302
 segmentation 227
 self-organizing maps (SOM) 32, 199, 201, 325, 327
 semantic Web 322
 semi-structured documents 439
 seq 112
 sequential-pattern mining 445
 shared memory 114
 shared memory multiprocessors (SMP) 108
 shared or integrated systems topologies 356
 shared tree (ST) 127
 shared-memory multiprocessors 107
 similarity search 445
 Simpson's paradox 13
 simulation techniques 260
 single driver single cost pool case 255
 site modification 440
 social network analysis 443
 software engineering 422
 software reuse 111
 soil fertility mapping 33
 spatial and geographic data 444
 spatial clustering 108, 128
 spatial data cubes 444
 spatial OLAP 444
 spatial queries 129
 spatial warehouses 444
 specialist-moderator (SM) 37
 specificity 167
 spider 304
 statistical inference 2
 statistical models 1
 streams 112
 strength 146
 strength factor 167

- structural organization of information 279
- structured parallel languages 107
- structured parallel version 106
- subjectivity 265
- summarization 231
- supervised learning networks 312
- support 146, 167
- syntactic objects 280
- system improvement 439
- system transparency 109

T

- task expansion policy 126
- task parallel 112
- task selection policy 127
- templates 111
- temporal naïve Bayesian network 38
- test set 122
- test statistic 5
- text categorization 200
- text clustering 440
- text documents 199
- text markups 442
- thematic hierarchies 199
- theory of rough sets (RST) 143
- threshold boundaries 180
- time series analysis 225
- time series learning architectures 45
- time-series and sequence-based data 444
- time-series data 253
- Titanic dataset 261
- total cost of ownership (TCO) 223
- training Bayesian belief networks 273
- training data 273
- training set 122
- transparency 29
- trigger 180

U

- u-lower approximation 155
- u-positive region 155, 156
- uncertainty management 6
- unstructured documents 438

- unsupervised competitive networks 313
- unsupervised learning 89, 324, 442
- upper approximation 152
- upper limit 155
- USA Patriot Act of 2001 414
- usage characterization 439

V

- value reduct 151
- variable precision rough set model (VPRSM) 144
- vector quantization 30
- vector space model 203
- vertical representation 115
- vertically partitioned data sets 441
- Video Privacy Protection Act 411
- virtual memory 224
- visual data mining (visual DM) 59, 442
- visualization 55

W

- Web caching 439
- Web-content mining 438
- Web-log mining 438
- Web mining 438
- Web-oriented query languages 438
- Web query systems 305
- Web querying 303
- Web search engines 303
- Web searching 303
- Web-structure mining 438
- Web-usage mining 438
- webot 303
- WebWatcher 305
- word cluster map (WCM) 201, 204
- word semantics 320
- World Wide Web 301
- wrapper approach 27
- wrappers 34

X

- XML 321

Z

- zero-frequency 268