

The Data Mining and Knowledge Discovery Handbook

Oded Maimon
Lior Rokach



Springer

Contents

Dedication	ii
Contributing Authors	xxi
Preface	xxxiii
Acknowledgments	xxxv
1	
Introduction to Knowledge Discovery in Databases	1
<i>Oded Maimon and Lior Rokach</i>	
1. The KDD Process	2
2. Taxonomy of Data Mining Methods	6
3. Data Mining within the Complete Decision Support System	8
4. KDD & DM Research Opportunities and Challenges	9
5. KDD & DM Trends	11
6. The Organization of the Handbook	12
7. References Principles	13
 Part I Preprocessing Methods	
2	
Data Cleansing	21
<i>Jonathan I. Maletic and Andrian Marcus</i>	
1. INTRODUCTION	21
2. DATA CLEANSING BACKGROUND	22
3. GENERAL METHODS FOR DATA CLEANSING	26
4. APPLYING DATA CLEANSING	27
5. CONCLUSIONS	33
References	33
3	
Handling Missing Attribute Values	37
<i>Jerzy W. Grzymala-Busse and Witold J. Grzymala-Busse</i>	
1. Introduction	37
2. Sequential Methods	39
3. Parallel Methods	48
4. Conclusions	53
References	53

4

Geometric Methods for Feature Extraction and Dimensional Reduction	59
--	----

Christopher J.C. Burges

1. Projective Methods	61
2. Manifold Modeling	73
3. Pulling the Threads Together	86

Acknowledgments	88
-----------------	----

References	89
------------	----

5

Dimension Reduction and Feature Selection	93
---	----

Barak Chizi and Oded Maimon

1. Introduction	93
2. Feature Selection Techniques	96
3. Variable Selection	106

References	109
------------	-----

6

Discretization Methods	113
------------------------	-----

Ying Yang, Geoffrey I. Webb and Xindong Wu

1. Terminology	114
2. Taxonomy	116
3. Typical methods	118

4. Discretization and the learning context	125
--	-----

5. Summary	127
------------	-----

References	128
------------	-----

7

Outlier Detection	131
-------------------	-----

Irad Ben-Gal

1. Introduction: Motivation, Definitions and Applications	131
2. Taxonomy of Outlier Detection Methods	132
3. Univariate Statistical Methods	133

4. Multivariate Outlier Detection	137
-----------------------------------	-----

5. Comparison of Outlier Detection Methods	141
--	-----

References	142
------------	-----

Part II Supervised Methods

8

Introduction to Supervised Methods	149
------------------------------------	-----

Oded Maimon and Lior Rokach

1. Introduction	149
-----------------	-----

2. Training Set	150
-----------------	-----

3. Definition of the Classification Problem	151
---	-----

4. Induction Algorithms	152
-------------------------	-----

5. Performance Evaluation	152
---------------------------	-----

6. Scalability to Large Datasets	158
----------------------------------	-----

7.	The “Curse of Dimensionality”	159
8.	Classification Problem Extensions	161
	References	162
9		
	Decision Trees	165
	<i>Lior Rokach and Oded Maimon</i>	
1.	Decision Trees	165
2.	Algorithmic Framework for Decision Trees	167
3.	Univariate Splitting Criteria	168
4.	Multivariate Splitting Criteria	174
5.	Stopping Criteria	174
6.	Pruning Methods	175
7.	Other Issues	179
8.	Decision Trees Inducers	181
9.	Advantages and Disadvantages of Decision Trees	183
10.	Decision Tree Extensions	185
	References	187
10		
	Bayesian Networks	193
	<i>Paola Sebastiani, Maria M. Abad and Marco F. Ramoni</i>	
1.	Introduction	193
2.	Representation	195
3.	Reasoning	198
4.	Learning	200
5.	Bayesian Networks in Data Mining	211
6.	Data Mining Applications	218
7.	Conclusions and Future Research Directions	223
	Acknowledgments	226
	References	226
11		
	Data Mining within a Regression Framework	231
	<i>Richard A. Berk</i>	
1.	Introduction	231
2.	Some Definitions	232
3.	Regression Splines	234
4.	Smoothing Splines	236
5.	Locally Weighted Regression as a Smoother	238
6.	Smoothers for Multiple Predictors	239
7.	Recursive Partitioning	242
8.	Conclusions	252
	Acknowledgments	252
	References	253

Support Vector Machines	257
<i>Armin Shmilovici</i>	
1. Introduction	257
2. Hyperplane Classifiers	259
3. Non-Separable SVM Models	264
4. Implementation Issues with SVM	269
5. Extensions and Application	272
6. Conclusion	273
References	273

Rule Induction	277
<i>Jerzy W. Grzymala-Busse</i>	
1. Introduction	277
2. Types of Rules	279
3. Rule Induction Algorithms	281
4. Classification Systems	291
5. Validation	292
6. Advanced Methodology	293
References	293

Part III Unsupervised Methods

Visualization and Data Mining for High Dimensional Datasets	297
<i>Alfred Inselberg</i>	
1. Do it in Parallel!	298
2. Visual Data Mining – A Case Study	304
3. Visual and Computational Models	316
References	318

Clustering Methods	321
<i>Lior Rokach and Oded Maimon</i>	
1. Introduction	321
2. Distance Measures	322
3. Similarity Functions	325
4. Evaluation Criteria Measures	326
5. Clustering Methods	330
6. Clustering Large Data Sets	342
7. Determining the Number of Clusters	346
References	349

Association Rules	353
<i>Frank Höppner</i>	
1. Introduction	353
2. Association Rule Mining	356

3.	Application to Other Types of Data	362
4.	Extensions of the Basic Framework	364
5.	Conclusions	372
	References	373
17		
	Frequent Set Mining	377
	<i>Bart Goethals</i>	
1.	Problem Description	378
2.	Apriori	381
3.	Eclat	384
4.	Optimizations	386
5.	Concise representations	388
6.	Theoretical Aspects	391
7.	Further Reading	392
	References	393
18		
	Constraint-based Data Mining	399
	<i>Jean-Francois Boulcaut and Baptiste Jeudy</i>	
1.	Motivations	399
2.	Background and Notations	402
3.	Solving Anti-Monotonic Constraints	404
4.	Introducing non Anti-Monotonic Constraints	406
5.	Conclusion	413
	References	414
19		
	Link Analysis	417
	<i>Steve Donoho</i>	
1.	Introduction	417
2.	Social Network Analysis	419
3.	Search Engines	422
4.	Viral Marketing	424
5.	Law Enforcement & Fraud Detection	426
6.	Combining with Traditional Methods	428
7.	Summary	430
	References	430
	 Part IV Soft Computing Methods	
20		
	Evolutionary Algorithms for Data Mining	435
	<i>Alex A. Freitas</i>	
1.	Introduction	435
2.	An Overview of Evolutionary Algorithms	436
3.	Evolutionary Algorithms for Discovering Classification Rules	442
4.	Evolutionary Algorithms for Clustering	447

5.	Evolutionary Algorithms for Data Preprocessing	450
6.	Multi-Objective Optimization with Evolutionary Algorithms	456
7.	Conclusions	459
	References	461
21		
	Reinforcement-Learning: an Overview from a Data Mining Perspective	469
	<i>Shahar Cohen and Oded Maimon</i>	
1.	Introduction	469
2.	The Reinforcement-Learning Model	470
3.	Reinforcement-Learning Algorithms	472
4.	Extensions to Basic Model and Algorithms	476
5.	Applications of Reinforcement-Learning	478
6.	Reinforcement-Learning and Data-Mining	479
7.	An Instructive Example	480
	References	485
22		
	Neural Networks	487
	<i>Peter G. Zhang</i>	
1.	Introduction	487
2.	A Brief History	488
3.	Neural Network Models	490
4.	Data Mining Applications	506
5.	Conclusions	508
	References	508
23		
	On the use of Fuzzy Logic in Data Mining	517
	<i>Joseph Komek and Moti Schneider</i>	
1.	Introduction	517
2.	Fuzzy Sets and Fuzzy Logic	518
3.	Soft Regression	522
4.	Fuzzy Association Rules	525
5.	Conclusions	532
	References	532
24		
	Granular Computing and Rough Sets	535
	<i>Tsau Young ('T. Y.') Lin and Churn-Jung Liao</i>	
1.	Introduction	535
2.	Naive Model for Problem Solving	536
3.	A Geometric Models of Information Granulations	538
4.	Information Granulations/Partitions	540
5.	Non-partition Application - Chinese Wall Security Policy Model	541
6.	Knowledge Representations	543
7.	Topological Concept Hierarchy Lattices/Trees	549
8.	Knowledge Processing	553

9.	Information Integration	556
10.	Conclusions	558
	References	558

Part V Supporting Methods

25

Statistical Methods for Data Mining	565
-------------------------------------	-----

Yoav Benjamini and Moshe Leshno

1.	Introduction	565
2.	Statistical Issues in DM	567
3.	Modeling Relationships using Regression Models	573
4.	False Discovery Rate (FDR) Control in Hypotheses Testing	578
5.	Model (Variables or Features) Selection using FDR Penalization in GLM	582
6.	Concluding Remarks	584
	References	585

26

Logics for Data Mining	589
------------------------	-----

Petr Hájek

1.	Generalized quantifiers	590
2.	Some important classes of quantifiers	593
3.	Some comments and conclusion	598

Acknowledgments	599
-----------------	-----

References	599
------------	-----

27

Wavelet Methods in Data Mining	603
--------------------------------	-----

Tao Li, Sheng Ma and Mitsunori Ogiwara

1.	Introduction	604
2.	A Framework for Data Mining Process	604
3.	Wavelet Background	605
4.	Data Management	610
5.	Preprocessing	611
6.	Core Mining Process	614
7.	Conclusion	622
	References	623

28

Fractal Mining	627
----------------	-----

Daniel Barbara and Ping Chen

1.	Introduction	628
2.	Fractal Dimension	629
3.	Clustering Using the Fractal Dimension	633
4.	Projected Fractal Clustering	641
5.	Tracking Clusters	642
6.	Conclusions	645

References	645
29	
Interestingness Measures	649
<i>Sigal Sahar</i>	
1. Definitions and Notations	650
2. Subjective Interestingness	651
3. Objective Interestingness	652
4. Impartial Interestingness	655
5. Concluding Remarks	656
References	657
30	
Quality Assessment Approaches in Data Mining	661
<i>Maria Halkidi and Michalis Vazirgiannis</i>	
1. Data Pre-processing and Quality Assessment	663
2. Evaluation of Classification Methods	664
3. Association Rules	671
4. Cluster Validity	675
References	694
31	
Data Mining Model Comparison	697
<i>Paolo Giudici</i>	
1. Data Mining and Statistics	697
2. Data Mining Model Comparison	698
3. Application to Credit Risk Management	704
4. Conclusions	712
References	714
32	
Data Mining Query Languages	715
<i>Jean-Francois Boulicaut and Cyrille Masson</i>	
1. The Need for Data Mining Query Languages	715
2. Supporting Association Rule Mining Processes	717
3. A Few Proposals for Association Rule Mining	719
4. Conclusion	725
References	726
Part VI Advanced Methods	
33	
Meta-Learning	731
<i>Ricardo Vilalta, Christophe Giraud-Carrier and Pavel Brazdil</i>	
1. Introduction	731
2. A Meta-Learning Architecture	733
3. Techniques in Meta-Learning	737
4. Tools and Applications	743

<i>Contents</i>	xiii
5. Future Directions and Conclusions	743
References	744
34	
Bias vs Variance Decomposition for Regression and Classification	749
<i>Pierre Geurts</i>	
1. Introduction	749
2. Bias/Variance Decompositions	751
3. Estimation of Bias and Variance	758
4. Experiments and Applications	760
5. Discussion	762
References	762
35	
Mining with Rare Cases	765
<i>Gary M. Weiss</i>	
1. Introduction	765
2. Why Rare Cases are Problematic	767
3. Techniques for Handling Rare Cases	770
4. Conclusion	774
References	775
36	
Mining Data Streams	777
<i>Haixun Wang, Philip S. Yu and Jiawei Han</i>	
1. Introduction	778
2. The Data Expiration Problem	779
3. Classifier Ensemble for Drifting Concepts	781
4. Experiments	783
5. Discussion and Related Work	789
References	790
37	
Mining High-Dimensional Data	793
<i>Wei Wang and Jiong Yang</i>	
1. Introduction	793
2. Challenges	794
3. Frequent Pattern	794
4. Clustering	795
5. Classification	797
References	798
38	
Text Mining and Information Extraction	801
<i>Moty Ben-Dov and Ronen Feldman</i>	
1. Introduction	801
2. Text Mining vs. Text Retrieval	803
3. Task-Oriented Approaches vs. Formal Frameworks	804
4. Task-Oriented Approaches	804

5.	Formal Frameworks And Algorithm-Based Techniques	808
6.	Hybrid Approaches - TEG	814
7.	Text Mining – Visualization and Analytics	815
	References	820
39		
	Spatial Data Mining	833
	<i>Shashi Shekhar, Pusheng Zhang and Yan Huang</i>	
1.	Introduction	833
2.	Spatial Data	834
3.	Spatial Outliers	837
4.	Spatial Co-location Rules	841
5.	Predictive Models	844
6.	Spatial Clusters	848
7.	Summary	849
	Acknowledgments	849
	References	850
40		
	Data Mining for Imbalanced Datasets: An Overview	853
	<i>Nitesh V. Chawla</i>	
1.	Introduction	853
2.	Performance Measure	854
3.	Sampling Strategies	858
4.	Ensemble-based Methods	860
5.	Discussion	862
	References	863
41		
	Relational Data Mining	869
	<i>Sašo Džeroski</i>	
1.	In a Nutshell	869
2.	Inductive logic programming	874
3.	Relational Association Rules	884
4.	Relational Decision Trees	889
5.	RDM Literature and Internet Resources	894
	References	895
42		
	Web Mining	899
	<i>Johannes Fürnkranz</i>	
1.	Introduction	899
2.	Graph Properties of the Web	900
3.	Web Search	902
4.	Text Classification	904
5.	Hypertext Classification	905
6.	Information Extraction and Wrapper Induction	907
7.	The Semantic Web	908

8.	Web Usage Mining	909
9.	Collaborative Filtering	910
10.	Conclusion	911
	References	911
43		
	A Review of Web Document Clustering Approaches	921
	<i>Nora Oikonomakou and Michalis Vazirgiannis</i>	
1.	Introduction	922
2.	Motivation for Document Clustering	922
3.	Web Document Clustering Approaches	924
4.	Comparison	935
5.	Conclusions and Open Issues	937
	References	937
44		
	Causal Discovery	945
	<i>Hong Yao, Cory J. Butz, and Howard J. Hamilton</i>	
1.	Introduction	945
2.	Background Knowledge	946
3.	Theoretical Foundation	949
4.	Learning a DAG of CN by FDs	950
5.	Experimental Results	953
6.	Conclusion	953
	References	954
45		
	Ensemble Methods For Classifiers	957
	<i>Lior Rokach</i>	
1.	Introduction	957
2.	Sequential Methodology	958
3.	Concurrent Methodology	964
4.	Combining Classifiers	966
5.	Ensemble Diversity	973
6.	Ensemble Size	974
7.	Cluster Ensemble	976
	References	977
46		
	Decomposition Methodology for Knowledge Discovery and Data Mining	981
	<i>Oded Maimon and Lior Rokach</i>	
1.	Introduction	981
2.	Decomposition Advantages	984
3.	The Elementary Decomposition Methodology	986
4.	The Decomposer's Characteristics	991
5.	The Relation to Other Methodologies	996
6.	Summary	999

References	999
47	
Information Fusion	1005
<i>Vicenc Torra</i>	
1. Introduction	1005
2. Preprocessing Data	1006
3. Building Data Models	1009
4. Information Extraction	1012
5. Conclusions	1012
Acknowledgments	1012
References	1013
48	
Parallel And Grid-Based Data Mining	1017
<i>Antonio Congiusta, Domenico Talia and Paolo Trunfio</i>	
1. Introduction	1018
2. Parallel Data Mining	1019
3. Grid-Based Data Mining	1027
4. The Knowledge Grid	1033
5. Summary	1038
References	1039
49	
Collaborative Data Mining	1043
<i>Steve Moyle</i>	
1. Introduction	1043
2. Remote Collaboration	1044
3. The Data Mining Process	1047
4. Collaborative Data Mining Guidelines	1048
5. Discussion	1052
6. Conclusions	1053
References	1054
50	
Organizational Data Mining	1057
<i>Hamid R. Nemati and Christopher D. Barko</i>	
1. Introduction	1058
2. Organizational Data Mining	1059
3. ODM versus Data Mining	1060
4. Ongoing ODM Research	1062
5. ODM Advantages	1062
6. ODM Evolution	1063
7. Summary	1066
References	1066
51	
Mining Time Series Data	1069

Chotirat Ann Ratanamahatana, Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh, Michail Vlachos and Gautam Das

1.	Introduction	1069
2.	Time Series Similarity Measures	1071
3.	Time Series Data Mining	1077
4.	Time Series Representations	1088
5.	Summary	1098
	References	1098

Part VII Applications

52		
Data Mining in Medicine		1107

Nada Lavrač and Blaž Zupan

1.	Introduction	1107
2.	Symbolic Classification Methods	1109
3.	Subsymbolic Classification Methods	1120
4.	Other Methods Supporting Medical Knowledge Discovery	1126
5.	Conclusions	1129

Acknowledgments		1129
-----------------	--	------

References		1129
------------	--	------

53		
Learning Information Patterns in Biological Databases		1139

Gautam B. Singh

1.	Background	1139
2.	Learning Stochastic Pattern Models	1141
3.	Searching for Meta-Patterns	1148
4.	Conclusions	1156
	References	1156

54		
Data Mining for Selection of Manufacturing Processes		1159

Bruno Agard and Andrew Kusiak

1.	Introduction	1159
2.	Data Mining in Engineering	1160
3.	Selection of Manufacturing Process with a Data Mining Approach	1161
4.	Conclusion	1165
	References	1166

55		
Data Mining of Design Products and Processes		1167

Yoram Reich

1.	Introduction	1167
2.	Product Design Process	1169
3.	Product Portfolio Management	1171
4.	Conceptual Design	1172

5.	Detailed Design	1175
6.	Business and Manufacturing Process Planning	1177
7.	Text Mining	1178
8.	Observations and Future Advancements	1180
9.	Epilogue	1182
	References	1183
56		
	Data Mining in Telecommunications	1189
	<i>Gary M. Weiss</i>	
1.	Introduction	1189
2.	Types of Telecommunication Data	1190
3.	Data Mining Applications	1194
4.	Conclusion	1199
	References	1200
57		
	Data Mining for Financial Applications	1203
	<i>Boris Kovalerchuk and Evgenii Vityaev</i>	
1.	Introduction: Financial Tasks	1203
2.	Specifics of Data Mining in Finance	1205
3.	Aspects of Data Mining Methodology in Finance	1210
4.	Data Mining Models and Practice in Finance	1214
5.	Conclusion	1219
	References	1221
58		
	Data Mining for Intrusion Detection	1225
	<i>Anoop Singhal and Sushil Jajodia</i>	
1.	Introduction	1225
2.	Data Mining Basics	1226
3.	Data Mining Meets Intrusion Detection	1228
4.	Conclusions and Future Research Directions	1235
	References	1236
59		
	Data Mining For Software Testing	1239
	<i>Mark Last</i>	
1.	Introduction	1239
2.	Mining Software Metrics Databases	1241
3.	Interaction-Pattern Discovery in System Usage Data	1242
4.	Using Data Mining in Functional Testing	1243
5.	Summary	1246
	Acknowledgments	1247
	References	1247
60		
	Data Mining for CRM	1249

Kurt Thearling

1. What is CRM? 1249
2. Data Mining and Campaign Management 1251
3. An Example: Customer Acquisition 1252

61

Data Mining for Target Marketing 1261

Nissan Levin and Jacob Zahavi

1. Introduction 1261
2. Modeling Process 1263
3. Evaluation Metrics 1265
4. Segmentation Methods 1268
5. Predictive Modeling 1275
6. In-Market Timing 1281
7. Pitfalls of Targeting 1285
8. Conclusions 1297
- References 1299

Part VIII Software

62

Weka 1305

Eibe Frank, Mark Hall, Geoffrey Holmes, Richard Kirkby, Bernhard Pfahringer and Ian H. Witten and Len Trigg

1. Introduction 1305
- References 1313

63

Oracle Data Mining 1315

Tamayo P., C. Berger, M. Campos, J. Yarmus, B. Milenova, A. Mozes, M. Taft, M. Hornick, R. Krishnan, S. Thomas, M. Kelly, D. Mukhin, B. Haberstroh, S. Stephens, and J. Myczkowski

1. Introduction 1315
2. The Mining-in-the-Database Paradigm 1317
3. ODM Functionality and Algorithms 1319
4. Text and Spatial Mining 1324
5. ODM Examples 1325
6. Conclusions 1327
- References 1328

64

Building Data Mining Solutions with OLE DB for DM and XML for Analysis 1331

Zhaohui Tang, Jamie MacLennan and Pyunghul (Peter) Kim

1. Introduction 1331
2. OLE DB for Data Mining 1332
3. Data Mining in SQL Server 2000 1336
4. Building Data Mining Application using OLE DB for Data Mining 1338
5. XML for Analysis 1340

6.	Conclusion	1343
	References	1344
65		
	LEERS—A Data Mining System	1347
	<i>Jerzy W. Grzymala-Busse</i>	
1.	Introduction	1347
2.	Input Data	1348
3.	Rule Sets	1348
4.	Main Features	1349
5.	Final Remarks	1350
	References	1350
66		
	GainSmarts Data Mining System for Marketing	1353
	<i>Nissan Levin and Jacob Zahavi</i>	
1.	Introduction	1353
2.	Accessing GainSmarts	1354
3.	Setting Up the Data for Modeling	1355
4.	GainSmarts Modules	1355
5.	Knowledge Evaluation	1360
6.	Reporting	1360
7.	Software Characteristics	1363
8.	Applications	1363
	References	1363
67		
	WizSoft's WizWhy	1365
	<i>Abraham Meidan</i>	
1.	Introduction	1365
2.	If-Then Rules	1366
3.	If-and-Only-If Rules	1366
4.	Data Summarization	1367
5.	Interesting Phenomena	1367
6.	Classifications	1368
7.	Data Auditing	1368
8.	WizWhy vs. other Data Mining Methods	1369
	References	1369
68		
	DataEngine	1371
	<i>Joseph Komem and Moti Schneider</i>	
1.	Overview	1371
2.	Intelligent Technologies for Modeling and Control	1372
3.	Work with DataEngine	1374
	References	1377
Index		
		1379