

RAPID MINER

Data Mining Use Cases and Business Analytics Applications

Edited by

Markus Hofmann

**Institute of Technology
Blanchardstown, Dublin, Ireland**

Ralf Klinkenberg

**Rapid-I / RapidMiner
Dortmund, Germany**



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business

A CHAPMAN & HALL BOOK

Contents

I	Introduction to Data Mining and RapidMiner	1
1	What This Book is About and What It is Not	3
	<i>Ingo Mierswa</i>	
1.1	Introduction	3
1.2	Coincidence or Not?	4
1.3	Applications of Data Mining	7
1.3.1	Financial Services	7
1.3.2	Retail and Consumer Products	8
1.3.3	Telecommunications and Media	9
1.3.4	Manufacturing, Construction, and Electronics	10
1.4	Fundamental Terms	11
1.4.1	Attributes and Target Attributes	11
1.4.2	Concepts and Examples	13
1.4.3	Attribute Roles	14
1.4.4	Value Types	14
1.4.5	Data and Meta Data	15
1.4.6	Modeling	16
2	Getting Used to RapidMiner	19
	<i>Ingo Mierswa</i>	
2.1	Introduction	19
2.2	First Start	19
2.3	Design Perspective	21
2.4	Building a First Process	23
2.4.1	Loading Data	24
2.4.2	Creating a Predictive Model	25
2.4.3	Executing a Process	28
2.4.4	Looking at Results	29
II	Basic Classification Use Cases for Credit Approval and in Education	31
3	k-Nearest Neighbor Classification I	33
	<i>M. Fareed Akhtar</i>	
3.1	Introduction	33
3.2	Algorithm	34
3.3	The k-NN Operator in RapidMiner	34
3.4	Dataset	35
3.4.1	Teacher Assistant Evaluation Dataset	35
3.4.2	Basic Information	35
3.4.3	Examples	35

3.4.4	Attributes	35
3.5	Operators in This Use Case	36
3.5.1	Read URL Operator	36
3.5.2	Rename Operator	36
3.5.3	Numerical to Binominal Operator	37
3.5.4	Numerical to Polynominal Operator	37
3.5.5	Set Role Operator	37
3.5.6	Split Validation Operator	37
3.5.7	Apply Model Operator	38
3.5.8	Performance Operator	38
3.6	Use Case	38
3.6.1	Data Import	39
3.6.2	Pre-processing	39
3.6.3	Renaming Attributes	40
3.6.4	Changing the Type of Attributes	40
3.6.5	Changing the Role of Attributes	41
3.6.6	Model Training, Testing, and Performance Evaluation	41
4	k-Nearest Neighbor Classification II	45
	<i>M. Fareed Akhtar</i>	
4.1	Introduction	45
4.2	Dataset	45
4.3	Operators Used in This Use Case	46
4.3.1	Read CSV Operator	46
4.3.2	Principal Component Analysis Operator	47
4.3.3	Split Data Operator	48
4.3.4	Performance (Classification) Operator	48
4.4	Data Import	48
4.5	Pre-processing	50
4.5.1	Principal Component Analysis	50
4.6	Model Training, Testing, and Performance Evaluation	50
4.6.1	Training the Model	51
4.6.2	Testing the Model	51
4.6.3	Performance Evaluation	51
5	Naïve Bayes Classification I	53
	<i>M. Fareed Akhtar</i>	
5.1	Introduction	53
5.2	Dataset	54
5.2.1	Credit Approval Dataset	54
5.2.2	Examples	54
5.2.3	Attributes	55
5.3	Operators in This Use Case	56
5.3.1	Rename by Replacing Operator	56
5.3.2	Filter Examples Operator	56
5.3.3	Discretize by Binning Operator	56
5.3.4	X-Validation Operator	57
5.3.5	Performance (Binominal Classification) Operator	57
5.4	Use Case	57
5.4.1	Data Import	58
5.4.2	Pre-processing	58

5.4.3	Model Training, Testing, and Performance Evaluation	61
6	Naïve Bayes Classification II	65
	<i>M. Fareed Akhtar</i>	
6.1	Dataset	65
6.1.1	Nursery Dataset	65
6.1.2	Basic Information	65
6.1.3	Examples	66
6.1.4	Attributes	66
6.2	Operators in this Use Case	67
6.2.1	Read Excel Operator	67
6.2.2	Select Attributes Operator	67
6.3	Use Case	67
6.3.1	Data Import	68
6.3.2	Pre-processing	69
6.3.3	Model Training, Testing, and Performance Evaluation	69
6.3.4	A Deeper Look into the Naïve Bayes Algorithm	71
III	Marketing, Cross-Selling, and Recommender System Use Cases	75
7	Who Wants My Product? Affinity-Based Marketing	77
	<i>Euler Timm</i>	
7.1	Introduction	77
7.2	Business Understanding	78
7.3	Data Understanding	79
7.4	Data Preparation	81
7.4.1	Assembling the Data	82
7.4.2	Preparing for Data Mining	86
7.5	Modelling and Evaluation	87
7.5.1	Continuous Evaluation and Cross Validation	87
7.5.2	Class Imbalance	88
7.5.3	Simple Model Evaluation	89
7.5.4	Confidence Values, ROC, and Lift Charts	90
7.5.5	Trying Different Models	92
7.6	Deployment	93
7.7	Conclusions	94
8	Basic Association Rule Mining in RapidMiner	97
	<i>Matthew A. North</i>	
8.1	Data Mining Case Study	97
9	Constructing Recommender Systems in RapidMiner	119
	<i>Matej Mihelčič, Matko Bošnjak, Nino Antulov-Fantulin, and Tomislav Šmuc</i>	
9.1	Introduction	120
9.2	The Recommender Extension	121
9.2.1	Recommendation Operators	121
9.2.2	Data Format	122
9.2.3	Performance Measures	124
9.3	The VideoLectures.net Dataset	126
9.4	Collaborative-based Systems	127

9.4.1	Neighbourhood-based Recommender Systems	127
9.4.2	Factorization-based Recommender Systems	128
9.4.3	Collaborative Recommender Workflows	130
9.4.4	Iterative Online Updates	131
9.5	Content-based Recommendation	132
9.5.1	Attribute-based Content Recommendation	133
9.5.2	Similarity-based Content Recommendation	134
9.6	Hybrid Recommender Systems	135
9.7	Providing RapidMiner Recommender System Workflows as Web Services Using RapidAnalytics	138
9.7.1	Simple Recommender System Web Service	138
9.7.2	Guidelines for Optimizing Workflows for Service Usage	139
9.8	Summary	141
10	Recommender System for Selection of the Right Study Program for Higher Education Students	145
	<i>Milan Vukićević, Miloš Jovanović, Boris Delibašić, and Milija Suknović</i>	
10.1	Introduction	146
10.2	Literature Review	146
10.3	Automatic Classification of Students using RapidMiner	147
10.3.1	Data	147
10.3.2	Processes	147
10.3.2.1	Simple Evaluation Process	150
10.3.2.2	Complex Process (with Feature Selection)	152
10.4	Results	154
10.5	Conclusion	155
IV	Clustering in Medical and Educational Domains	157
11	Visualising Clustering Validity Measures	159
	<i>Andrew Chisholm</i>	
11.1	Overview	160
11.2	Clustering	160
11.2.1	A Brief Explanation of k-Means.	161
11.3	Cluster Validity Measures	161
11.3.1	Internal Validity Measures	161
11.3.2	External Validity Measures	162
11.3.3	Relative Validity Measures	163
11.4	The Data	163
11.4.1	Artificial Data	164
11.4.2	<i>E-coli</i> Data	164
11.5	Setup	165
11.5.1	Download and Install R Extension	166
11.5.2	Processes and Data	166
11.6	The Process in Detail	167
11.6.1	Import Data (A)	168
11.6.2	Generate Clusters (B)	169
11.6.3	Generate Ground Truth Validity Measures (C)	170
11.6.4	Generate External Validity Measures (D)	172
11.6.5	Generate Internal Validity Measures (E)	173
11.6.6	Output Results (F)	174

11.7	Running the Process and Displaying Results	175
11.8	Results and Interpretation	176
11.8.1	Artificial Data	176
11.8.2	<i>E-coli</i> Data	178
11.9	Conclusion	181
12	Grouping Higher Education Students with RapidMiner	185
	<i>Milan Vukićević, Miloš Jovanović, Boris Delibašić, and Milija Suknović</i>	
12.1	Introduction	185
12.2	Related Work	186
12.3	Using RapidMiner for Clustering Higher Education Students	186
12.3.1	Data	187
12.3.2	Process for Automatic Evaluation of Clustering Algorithms	187
12.3.3	Results and Discussion	191
12.4	Conclusion	193
V	Text Mining: Spam Detection, Language Detection, and Customer Feedback Analysis	197
13	Detecting Text Message Spam	199
	<i>Neil McGuigan</i>	
13.1	Overview	200
13.2	Applying This Technique in Other Domains	200
13.3	Installing the Text Processing Extension	200
13.4	Getting the Data	201
13.5	Loading the Text	201
13.5.1	Data Import Wizard Step 1	201
13.5.2	Data Import Wizard Step 2	202
13.5.3	Data Import Wizard Step 3	202
13.5.4	Data Import Wizard Step 4	202
13.5.5	Step 5	202
13.6	Examining the Text	203
13.6.1	Tokenizing the Document	203
13.6.2	Creating the Word List and Word Vector	204
13.6.3	Examining the Word Vector	204
13.7	Processing the Text for Classification	205
13.7.1	Text Processing Concepts	206
13.8	The Naïve Bayes Algorithm	207
13.8.1	How It Works	207
13.9	Classifying the Data as Spam or Ham	208
13.10	Validating the Model	208
13.11	Applying the Model to New Data	209
13.11.1	Running the Model on New Data	210
13.12	Improvements	210
13.13	Summary	211
14	Robust Language Identification with RapidMiner: A Text Mining Use Case	213
	<i>Matko Bošnjak, Eduarda Mendes Rodrigues, and Luis Sarmento</i>	
14.1	Introduction	214
14.2	The Problem of Language Identification	215

14.3	Text Representation	217
14.3.1	Encoding	217
14.3.2	Token-based Representation	218
14.3.3	Character-Based Representation	219
14.3.4	Bag-of-Words Representation	219
14.4	Classification Models	220
14.5	Implementation in RapidMiner	221
14.5.1	Datasets	221
14.5.2	Importing Data	223
14.5.3	Frequent Words Model	225
14.5.4	Character n-Grams Model	229
14.5.5	Similarity-based Approach	232
14.6	Application	234
14.6.1	RapidAnalytics	234
14.6.2	Web Page Language Identification	234
14.7	Summary	237
15	Text Mining with RapidMiner	241
	<i>Gurdal Ertek, Dilek Tapucu, and Inanc Arin</i>	
15.1	Introduction	242
15.1.1	Text Mining	242
15.1.2	Data Description	242
15.1.3	Running RapidMiner	242
15.1.4	RapidMiner Text Processing Extension Package	243
15.1.5	Installing Text Mining Extensions	243
15.2	Association Mining of Text Document Collection (Process01)	243
15.2.1	Importing Process01	243
15.2.2	Operators in Process01	243
15.2.3	Saving Process01	247
15.3	Clustering Text Documents (Process02)	248
15.3.1	Importing Process02	248
15.3.2	Operators in Process02	248
15.3.3	Saving Process02	250
15.4	Running Process01 and Analyzing the Results	250
15.4.1	Running Process01	250
15.4.2	Empty Results for Process01	252
15.4.3	Specifying the Source Data for Process01	252
15.4.4	Re-Running Process01	253
15.4.5	Process01 Results	253
15.4.6	Saving Process01 Results	257
15.5	Running Process02 and Analyzing the Results	257
15.5.1	Running Process02	257
15.5.2	Specifying the Source Data for Process02	257
15.5.3	Process02 Results	257

15.6	Conclusions	261
VI	Feature Selection and Classification in Astroparticle Physics and in Medical Domains	263
16	Application of RapidMiner in Neutrino Astronomy	265
	<i>Tim Ruhe, Katharina Morik, and Wolfgang Rhode</i>	
16.1	Protons, Photons, and Neutrinos	265
16.2	Neutrino Astronomy	267
16.3	Feature Selection	269
16.3.1	Installation of the Feature Selection Extension for RapidMiner . . .	269
16.3.2	Feature Selection Setup	270
16.3.3	Inner Process of the LOOP PARAMETERS Operator	271
16.3.4	Inner Operators of the WRAPPER X-VALIDATION	272
16.3.5	Settings of the LOOP PARAMETERS Operator	274
16.3.6	Feature Selection Stability	275
16.4	Event Selection Using a Random Forest	277
16.4.1	The Training Setup	278
16.4.2	The Random Forest in Greater Detail	280
16.4.3	The Random Forest Settings	281
16.4.4	The Testing Setup	282
16.5	Summary and Outlook	285
17	Medical Data Mining	289
	<i>Mertik Matej and Palfy Miroslav</i>	
17.1	Background	290
17.2	Description of Problem Domain: Two Medical Examples	291
17.2.1	Carpal Tunnel Syndrome	291
17.2.2	Diabetes	292
17.3	Data Mining Algorithms in Medicine	292
17.3.1	Predictive Data Mining	292
17.3.2	Descriptive Data Mining	293
17.3.3	Data Mining and Statistics: Hypothesis Testing	294
17.4	Knowledge Discovery Process in RapidMiner: Carpal Tunnel Syndrome .	295
17.4.1	Defining the Problem, Setting the Goals	295
17.4.2	Dataset Representation	295
17.4.3	Data Preparation	296
17.4.4	Modeling	298
17.4.5	Selecting Appropriate Methods for Classification	298
17.4.6	Results and Data Visualisation	303
17.4.7	Interpretation of the Results	303
17.4.8	Hypothesis Testing and Statistical Analysis	304
17.4.9	Results and Visualisation	308
17.5	Knowledge Discovery Process in RapidMiner: Diabetes	308
17.5.1	Problem Definition, Setting the Goals	309
17.5.2	Data Preparation	309
17.5.3	Modeling	310
17.5.4	Results and Data Visualization	312
17.5.5	Hypothesis Testing	313
17.6	Specifics in Medical Data Mining	316
17.7	Summary	316

VII	Molecular Structure- and Property-Activity Relationship Modeling in Biochemistry and Medicine	319
18	Using PaDEL to Calculate Molecular Properties and Chemoinformatic Models	321
	<i>Markus Muehlbacher and Johannes Kornhuber</i>	
18.1	Introduction	321
18.2	Molecular Structure Formats for Chemoinformatics	321
18.3	Installation of the PaDEL Extension for RapidMiner	322
18.4	Applications and Capabilities of the PaDEL Extension	323
18.5	Examples of Computer-aided Predictions	324
18.6	Calculation of Molecular Properties	325
18.7	Generation of a Linear Regression Model	325
18.8	Example Workflow	326
18.9	Summary	328
19	Chemoinformatics: Structure- and Property-activity Relationship Development	331
	<i>Markus Muehlbacher and Johannes Kornhuber</i>	
19.1	Introduction	331
19.2	Example Workflow	332
19.3	Importing the Example Set	332
19.4	Preprocessing of the Data	333
19.5	Feature Selection	334
19.6	Model Generation	335
19.7	Validation	337
19.8	Y-Randomization	338
19.9	Results	339
19.10	Conclusion/Summary	340
VIII	Image Mining: Feature Extraction, Segmentation, and Classification	345
20	Image Mining Extension for RapidMiner (Introductory)	347
	<i>Radim Burget, Václav Uher, and Jan Masek</i>	
20.1	Introduction	348
20.2	Image Reading/Writing	349
20.3	Conversion between Colour and Grayscale Images	352
20.4	Feature Extraction	353
	20.4.1 Local Level Feature Extraction	354
	20.4.2 Segment-Level Feature Extraction	356
	20.4.3 Global-Level Feature Extraction	358
20.5	Summary	359
21	Image Mining Extension for RapidMiner (Advanced)	363
	<i>Václav Uher and Radim Burget</i>	
21.1	Introduction	363
21.2	Image Classification	364
	21.2.1 Load Images and Assign Labels	364
	21.2.2 Global Feature Extraction	365
21.3	Pattern Detection	368

21.3.1	Process Creation	370
21.4	Image Segmentation and Feature Extraction	372
21.5	Summary	373
IX	Anomaly Detection, Instance Selection, and Prototype Construction	375
22	Instance Selection in RapidMiner	377
	<i>Marcin Blachnik and Mirosław Kordos</i>	
22.1	Introduction	377
22.2	Instance Selection and Prototype-Based Rule Extension	378
22.3	Instance Selection	379
22.3.1	Description of the Implemented Algorithms	381
22.3.2	Accelerating 1-NN Classification	384
22.3.3	Outlier Elimination and Noise Reduction	389
22.3.4	Advances in Instance Selection	392
22.4	Prototype Construction Methods	395
22.5	Mining Large Datasets	401
22.6	Summary	406
23	Anomaly Detection	409
	<i>Markus Goldstein</i>	
23.1	Introduction	410
23.2	Categorizing an Anomaly Detection Problem	412
23.2.1	Type of Anomaly Detection Problem (Pre-processing)	412
23.2.2	Local versus Global Problems	416
23.2.3	Availability of Labels	416
23.3	A Simple Artificial Unsupervised Anomaly Detection Example	417
23.4	Unsupervised Anomaly Detection Algorithms	419
23.4.1	k-NN Global Anomaly Score	419
23.4.2	Local Outlier Factor (LOF)	420
23.4.3	Connectivity-Based Outlier Factor (COF)	421
23.4.4	Influenced Outlierness (INFLO)	422
23.4.5	Local Outlier Probability (LoOP)	422
23.4.6	Local Correlation Integral (LOCI) and aLOCI	422
23.4.7	Cluster-Based Local Outlier Factor (CBLOF)	423
23.4.8	Local Density Cluster-Based Outlier Factor (LDCOF)	424
23.5	An Advanced Unsupervised Anomaly Detection Example	425
23.6	Semi-supervised Anomaly Detection	428
23.6.1	Using a One-Class Support Vector Machine (SVM)	428
23.6.2	Clustering and Distance Computations for Detecting Anomalies	430
23.7	Summary	433
X	Meta-Learning, Automated Learner Selection, Feature Selection, and Parameter Optimization	437
24	Using RapidMiner for Research: Experimental Evaluation of Learners	439
	<i>Jovanović Miloš, Vukićević Milan, Delibašić Boris, and Suknović Milija</i>	
24.1	Introduction	439
24.2	Research of Learning Algorithms	440
24.2.1	Sources of Variation and Control	440

- 24.2.2 Example of an Experimental Setup 441
- 24.3 Experimental Evaluation in RapidMiner 442
 - 24.3.1 Setting Up the Evaluation Scheme 442
 - 24.3.2 Looping Through a Collection of Datasets 443
 - 24.3.3 Looping Through a Collection of Learning Algorithms 445
 - 24.3.4 Logging and Visualizing the Results 445
 - 24.3.5 Statistical Analysis of the Results 447
 - 24.3.6 Exception Handling and Parallelization 449
 - 24.3.7 Setup for Meta-Learning 450
- 24.4 Conclusions 452

- Subject Index 455
- Operator Index 463