
Contents

Preface	xv
Contributors	xvii
1 Preliminaries and Overview	1
<i>Guozhu Dong and Huan Liu</i>	
1.1 Preliminaries	1
1.1.1 Features	1
1.1.2 Feature Engineering	3
1.1.3 Machine Learning and Data Analytic Tasks	3
1.2 Overview of the Chapters	4
1.3 Beyond this Book	7
1.3.1 Feature Engineering for Specific Data Types	8
1.3.2 Feature Engineering on Non-Data-Specific Topics	9
I Feature Engineering for Various Data Types	13
2 Feature Engineering for Text Data	15
<i>Chase Geigle, Qiaozhu Mei, and ChengXiang Zhai</i>	
2.1 Introduction	16
2.2 Overview of Text Representation	17
2.3 Text as Strings	18
2.4 Sequence of Words Representation	19
2.5 Bag of Words Representation	21
2.5.1 Term Weighting	22
2.5.2 Beyond Single Words	27
2.6 Structural Representation of Text	28
2.6.1 Semantic Structure Features	30
2.7 Latent Semantic Representation	31
2.7.1 Latent Semantic Analysis	31
2.7.2 Probabilistic Latent Semantic Analysis	33
2.7.3 Latent Dirichlet Allocation	35
2.8 Explicit Semantic Representation	37
2.9 Embeddings for Text Representation	37
2.9.1 Matrix Factorization for Word Embeddings	38
2.9.2 Neural Networks for Word Embeddings	40

2.9.3	Document Representations from Word Embeddings . .	41
2.10	Context-Sensitive Text Representation	42
2.11	Summary	45
3	Feature Extraction and Learning for Visual Data	55
	<i>Parag S. Chandakkar, Ragav Venkatesan, and Baoxin Li</i>	
3.1	Classical Visual Feature Representations	57
3.1.1	Color Features	57
3.1.2	Texture Features	61
3.1.3	Shape Features	63
3.2	Latent Feature Extraction	66
3.2.1	Principal Component Analysis	67
3.2.2	Kernel Principal Component Analysis	68
3.2.3	Multidimensional Scaling	69
3.2.4	Isomap	69
3.2.5	Laplacian Eigenmaps	70
3.3	Deep Image Features	71
3.3.1	Convolutional Neural Networks	72
3.3.1.1	The Dot-Product Layer	72
3.3.1.2	The Convolution Layer	73
3.3.2	CNN Architecture Design	75
3.3.3	Fine-Tuning Off-the-Shelf Neural Networks	76
3.3.4	Summary and Conclusions	79
4	Feature-Based Time-Series Analysis	87
	<i>Ben D. Fulcher</i>	
4.1	Introduction	87
4.1.1	The Time Series Data Type	87
4.1.2	Time-Series Characterization	89
4.1.3	Applications of Time-Series Analysis	90
4.2	Feature-Based Representations of Time Series	92
4.3	Global Features	95
4.3.1	Examples of Global Features	95
4.3.2	Massive Feature Vectors and Highly Comparative Time-Series Analysis	98
4.4	Subsequence Features	102
4.4.1	Interval Features	102
4.4.2	Shapelets	103
4.4.3	Pattern Dictionaries	105
4.5	Combining Time-Series Representations	106
4.6	Feature-Based Forecasting	108
4.7	Summary and Outlook	109

5	Feature Engineering for Data Streams	117
	<i>Yao Ma, Jiliang Tang, and Charu Aggarwal</i>	
5.1	Introduction	118
5.2	Streaming Settings	119
5.3	Linear Methods for Streaming Feature Construction	121
5.3.1	Principal Component Analysis for Data Streams	121
5.3.2	Linear Discriminant Analysis for Data Streams	123
5.4	Non-Linear Methods for Streaming Feature Construction	125
5.4.1	Locally Linear Embedding for Data Streams	125
5.4.2	Kernel Learning for Data Streams	126
5.4.3	Neural Networks for Data Streams	128
5.4.4	Discussion	132
5.5	Feature Selection for Data Streams with Streaming Features	132
5.5.1	The Grafting Algorithm	133
5.5.2	The Alpha-Investing Algorithm	133
5.5.3	The Online Streaming Feature Selection Algorithm	134
5.5.4	Unsupervised Streaming Feature Selection in Social Media	135
5.6	Feature Selection for Data Streams with Streaming Instances	135
5.6.1	Online Feature Selection	136
5.6.2	Unsupervised Feature Selection on Data Streams	136
5.7	Discussions and Challenges	136
5.7.1	Stability	137
5.7.2	Number of Features	137
5.7.3	Heterogeneous Streaming Data	137
6	Feature Generation and Feature Engineering for Sequences	145
	<i>Guozhu Dong, Lei Duan, Jyrki Nummenmaa, and Peng Zhang</i>	
6.1	Introduction	146
6.2	Basics on Sequence Data and Sequence Patterns	148
6.3	Approaches to Using Patterns in Sequence Features	149
6.4	Traditional Pattern-Based Sequence Features	150
6.5	Mined Sequence Patterns for Use in Sequence Features	151
6.5.1	Frequent Sequence Patterns	152
6.5.2	Closed Sequential Patterns	154
6.5.3	Gap Constraints for Sequence Patterns	155
6.5.4	Partial Order Patterns	156
6.5.5	Periodic Sequence Patterns	158
6.5.6	Distinguishing Sequence Patterns	158
6.5.7	Pattern Matching for Sequences	160
6.6	Factors for Selecting Sequence Patterns as Features	161
6.7	Sequence Features Not Defined by Patterns	161
6.8	Sequence Databases	162
6.9	Concluding Remarks	163

7	Feature Generation for Graphs and Networks	167
	<i>Yuan Yao, Hanghang Tong, Feng Xu, and Jian Lu</i>	
7.1	Introduction	168
7.2	Feature Types	168
7.3	Feature Generation	169
7.3.1	Basic Models	170
7.3.2	Extensions	175
7.3.3	Summary	179
7.4	Feature Usages	181
7.4.1	Multi-Label Classification	181
7.4.2	Link Prediction	181
7.4.3	Anomaly Detection	182
7.4.4	Visualization	182
7.5	Conclusions and Future Directions	183
7.6	Glossary	188
II	General Feature Engineering Techniques	189
8	Feature Selection and Evaluation	191
	<i>Yun Li and Tao Li</i>	
8.1	Introduction	191
8.2	Feature Selection Frameworks	192
8.2.1	Search-Based Feature Selection Framework	193
8.2.2	Correlation-Based Feature Selection Framework	194
8.3	Advanced Topics for Feature Selection	196
8.3.1	Stable Feature Selection	196
8.3.2	Sparsity-Based Feature Selection	199
8.3.3	Multi-Source Feature Selection	200
8.3.4	Distributed Feature Selection	203
8.3.5	Multi-View Feature Selection	204
8.3.6	Multi-Label Feature Selection	205
8.3.7	Online Feature Selection	206
8.3.8	Privacy-Preserving Feature Selection	208
8.3.9	Adversarial Feature Selection	210
8.4	Future Work and Conclusion	211
9	Automating Feature Engineering in Supervised Learning	221
	<i>Udayan Khurana</i>	
9.1	Introduction	222
9.1.1	Challenges in Performing Feature Engineering	224
9.2	Terminology and Problem Definition	225
9.3	A Few Simple Approaches	226
9.4	Hierarchical Exploration of Feature Transformations	227
9.4.1	Transformation Graph	228
9.4.2	Transformation Graph Exploration	229

9.5	Learning Optimal Traversal Policy	231
9.5.1	Feature Exploration through Reinforcement Learning	233
9.6	Finding Effective Features without Model Training	235
9.6.1	Learning to Predict Useful Transformations	237
9.7	Miscellaneous	239
9.7.1	Other Related Work	239
9.7.2	Research Opportunities	240
9.7.3	Resources	240
10	Pattern-Based Feature Generation	245
	<i>Yunzhe Jia, James Bailey, Ramamohanarao Kotagiri, and Christopher Leckie</i>	
10.1	Introduction	246
10.2	Preliminaries	247
10.2.1	Data and Patterns	247
10.2.2	Patterns for Non-Transactional Data	248
10.3	Framework of Pattern-Based Feature Generation	251
10.3.1	Pattern Mining	251
10.3.2	Pattern Selection	252
10.3.3	Feature Generation	253
10.4	Pattern Mining Algorithms	254
10.4.1	Frequent Pattern Mining	254
10.4.2	Contrast Pattern Mining	256
10.5	Pattern Selection Approaches	258
10.5.1	Post-Processing Pruning	258
10.5.2	In-processing Pruning	260
10.6	Pattern-Based Feature Generation	262
10.6.1	Unsupervised Mapping Functions	262
10.6.2	Supervised Mapping Functions	263
10.6.3	Feature Generation for Sequence Data and Graph Data	265
10.6.4	Comparison with Similar Techniques	265
10.7	Pattern-Based Feature Generation for Classification	266
10.7.1	Problem Statement	266
10.7.2	Direct Classification in the Pattern Space	267
10.7.3	Indirect Classification in the Pattern Space	268
10.7.4	Connection with Stacking Technique	269
10.8	Pattern-Based Feature Generation for Clustering	269
10.8.1	Clustering in the Pattern Space	269
10.8.2	Subspace Clustering	270
10.9	Conclusion	271
11	Deep Learning for Feature Representation	279
	<i>Suhang Wang and Huan Liu</i>	
11.1	Introduction	279
11.2	Restricted Boltzmann Machine	280

11.2.1	Deep Belief Networks and Deep Boltzmann Machine	281
11.2.2	RBM for Real-Valued Data	283
11.3	AutoEncoder	284
11.3.1	Sparse Autoencoder	286
11.3.2	Denoising Autoencoder	287
11.3.3	Stacked Autoencoder	287
11.4	Convolutional Neural Networks	288
11.4.1	Transfer Feature Learning of CNN	290
11.5	Word Embedding and Recurrent Neural Networks	291
11.5.1	Word Embedding	291
11.5.2	Recurrent Neural Networks	294
11.5.3	Gated Recurrent Unit	295
11.5.4	Long Short-Term Memory	296
11.6	Generative Adversarial Networks and Variational Autoencoder	296
11.6.1	Generative Adversarial Networks	297
11.6.2	Variational Autoencoder	298
11.7	Discussion and Further Readings	299

III Feature Engineering in Special Applications 309

12 Feature Engineering for Social Bot Detection 311

Onur Varol, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini

12.1	Introduction	312
12.2	Social Bot Detection	312
12.2.1	Holistic Approach	313
12.2.2	Pairwise Account Comparison	313
12.2.3	Egocentric Analysis	314
2.3	Online Bot Detection Framework	314
12.3.1	Feature Extraction	315
12.3.1.1	User-Based Features	316
12.3.1.2	Friend Features	316
12.3.1.3	Network Features	318
12.3.1.4	Content and Language Features	318
12.3.1.5	Sentiment Features	319
12.3.1.6	Temporal Features	320
12.3.2	Possible Directions for Feature Engineering	320
12.3.3	Feature Analysis	320
12.3.4	Feature Selection	323
12.3.4.1	Feature Classes	323
12.3.4.2	Top Individual Features	324
12.4	Conclusions	325
12.5	Glossary	334

13 Feature Generation and Engineering for Software Analytics 335*Xin Xia and David Lo*

13.1	Introduction	336
13.2	Features for Defect Prediction	337
13.2.1	File-level Defect Prediction	337
13.2.1.1	Code Features	338
13.2.1.2	Process Features	340
13.2.2	Just-in-time Defect Prediction	341
13.2.3	Prediction Models and Results	343
13.3	Features for Crash Release Prediction for Apps	343
13.3.1	Complexity Dimension	344
13.3.2	Time Dimension	345
13.3.3	Code Dimension	346
13.3.4	Diffusion Dimension	346
13.3.5	Commit Dimension	347
13.3.6	Text Dimension	347
13.3.7	Prediction Models and Results	348
13.4	Features from Mining Monthly Reports to Predict Developer Turnover	348
13.4.1	Working Hours	349
13.4.2	Task Report	349
13.4.3	Project	350
13.4.4	Prediction Models and Results	351
13.5	Summary	351

14 Feature Engineering for Twitter-Based Applications 359

*Sanjaya Wijeratne, Amit Sheth, Shreyansh Bhatt, Lakshika Balasuriya,
Hussein S. Al-Olimat, Manas Gaur, Amir Hossein Yazdavar,
Krishnaprasad Thirunarayan*

14.1	Introduction	359
14.2	Data Present in a Tweet	361
14.2.1	Tweet Text-Related Data	362
14.2.2	Twitter User-Related Data	363
14.2.3	Other Metadata	364
14.3	Common Types of Features Used in Twitter-Based Applications	364
14.3.1	Textual Features	365
14.3.2	Image and Video Features	368
14.3.3	Twitter Metadata-Related Features	369
14.3.4	Network Features	370
14.4	Twitter Feature Engineering in Selected Twitter-Based Studies	370
14.4.1	Twitter User Profile Classification	371
14.4.2	Assisting Coordination during Crisis Events	372
14.4.3	Location Extraction from Tweets	375
14.4.4	Studying the Mental Health Conditions of Depressed Twitter Users	377

14.4.5 Sentiment and Emotion Analysis on Twitter	379
14.5 Twitris: A Real-Time Social Media Analysis Platform	381
14.6 Conclusion	383
14.7 Acknowledgment	9

Index