Contents

Preface xi

Part 1. Symbolic Data 1

Chapter 1. Explanatory Tools for Machine Learning in the Symbolic Data Analysis Framework 3
Edwin DIDAY

1.1. Introduction 4
1.2. Introduction to Symbolic Data Analysis 6
1.2.1. What are complex data? 6
1.2.2. What are “classes” and “class of complex data”? 7
1.2.3. Which kind of class variability? 7
1.2.4. What are “symbolic variables” and “symbolic data tables”? 7
1.2.5. Symbolic Data Analysis (SDA) 9
1.3. Symbolic data tables from Dynamic Clustering Method and EM 10
1.3.1. The “dynamical clustering method” (DCM) 10
1.3.2. Examples of DCM applications 10
1.3.3. Clustering methods by mixture decomposition 12
1.3.4. Symbolic data tables from clustering 13
1.3.5. A general way to compare results of clustering methods by the “explanatory power” of their associated symbolic data table . . . 15
1.3.6. Quality criteria of classes and variables based on the cells of the symbolic data table containing intervals or inferred distributions . 15
1.4. Criteria for ranking individuals, classes and their bar chart descriptive symbolic variables 16
1.4.1. A theoretical framework for SDA 16
1.4.2. Characterization of a category and a class by a measure of discordance 18

1.4.3. Link between a characterization by the criteria W and the standard Tf-Idf 19
1.4.4. Ranking the individuals, the symbolic variables and the classes of a bar chart symbolic data table 21
1.5. Two directions of research 23
1.5.1. Parametrization of concordance and discordance criteria 23
1.5.2. Improving the explanatory power of any machine learning tool by a filtering process 25
1.6. Conclusion 27
1.7. References 28

Chapter 2. Likelihood in the Symbolic Context 31
Richard EMILION and Edwin DIDAY

2.1. Introduction 31
2.2. Probabilistic setting 32
2.2.1. Description variable and class variable 32
2.2.2. Conditional distributions 33
2.2.3. Symbolic variables 33
2.2.4. Examples 35
2.2.5. Probability measures on $(\mathbb{C}, \mathcal{C})$, likelihood 37
2.3. Parametric models for $p = 1$ 38
2.3.1. LDA model 38
2.3.2. BLS method 41
2.3.3. Interval-valued variables 42
2.3.4. Probability vectors and histogram-valued variables 42
2.4. Nonparametric estimation for $p = 1$ 45
2.4.1. Multihistograms and multivariate polygons 45
2.4.2. Dirichlet kernel mixtures 45
2.4.3. Dirichlet Process Mixture (DPM) 45
2.5. Density models for $p \geq 2$ 46
2.6. Conclusion 46
2.7. References 47

Chapter 3. Dimension Reduction and Visualization of Symbolic Interval-Valued Data Using Sliced Inverse Regression 49
Han-Ming WU, Chiun-How KAO and Chun-houh CHEN

3.1. Introduction 49
3.2. PCA for interval-valued data and the sliced inverse regression 51
3.2.1. PCA for interval-valued data 51
3.2.2. Classic SIR 52
3.3. SIR for interval-valued data 53
3.3.1. Quantification approaches 54
3.3.2. Distributional approaches 56

3.4. Projections and visualization in DR subspace 58
3.4.1. Linear combinations of intervals 58
3.4.2. The graphical representation of the projected intervals in the 2D DR subspace 59
3.5. Some computational issues 61
3.5.1. Standardization of interval-valued data 61
3.5.2. The slicing schemes for iSIR 62
3.5.3. The evaluation of DR components 62
3.6. Simulation studies 63
3.6.1. Scenario 1: aggregated data 63
3.6.2. Scenario 2: data based on interval arithmetic 63
3.6.3. Results 64
3.7. A real data example: face recognition data 65
3.8. Conclusion and discussion 73
3.9. References 74

Chapter 4. On the “Complexity” of Social Reality. Some Reflections About the Use of Symbolic Data Analysis in Social Sciences 79
Frédéric LEBARON

4.1. Introduction 79
4.2. Social sciences facing “complexity” 80
4.2.1. The total social fact, a designation of “complexity” in social sciences 80
4.2.2. Two families of answers 80
4.2.3. The contemporary deepening of the two approaches, “reductionist” and “encompassing” 81
4.2.4. Issues of scale and heterogeneity 82
4.3. Symbolic data analysis in the social sciences: an example 83
4.3.1. Symbolic data analysis 83
4.3.2. An exploratory case study on European data 83
4.3.3. A sociological interpretation 94
4.4. Conclusion 95
4.5. References 96

Part 2. Complex Data 99

Chapter 5. A Spatial Dependence Measure and Prediction of Georeferenced Data Streams Summarized by Histograms 101
Rosanna VERDE and Antonio BALZANELLA

5.1. Introduction 101
5.2. Processing setup 103
5.3. Main definitions 104
5.4. Online summarization of a data stream through CluStream for Histogram data 106
5.5. Spatial dependence monitoring: a variogram for histogram data 107
5.6. Ordinary kriging for histogram data 110

5.7. Experimental results on real data 112
5.8. Conclusion 116
5.9. References 116

Chapter 6. Incremental Calculation Framework for Complex Data . 119
Huiwen WANG, Yuan WEI and Siyang WANG

6.1. Introduction 119
6.2. Basic data 122
6.2.1. The basic data space 122
6.2.2. Sample covariance matrix 123
6.3. Incremental calculation of complex data 124
6.3.1. Transformation of complex data 124
6.3.2. Online decomposition of covariance matrix 125
6.3.3. Adopted algorithms 128
6.4. Simulation studies 131
6.4.1. Functional linear regression 131
6.4.2. Compositional PCA 133
6.5. Conclusion 135
6.6. Acknowledgment 135
6.7. References 135

Part 3. Network Data 139

Chapter 7. Recommender Systems and Attributed Networks 141
Françoise FOGELMAN-SOULIÉ, Lanxiang MEI, Jianyu ZHANG, Yiming LI, Wen GE, Yinglan LI and Qiaofei YE

7.1. Introduction 141
7.2. Recommender systems 142
7.2.1. Data used 143
7.2.2. Model-based collaborative filtering 145
7.2.3. Neighborhood-based collaborative filtering 145
7.2.4. Hybrid models 148
7.3. Social networks 150
7.3.1. Non-independence 150
7.3.2. Definition of a social network 150
7.3.3. Properties of social networks 151
7.3.4. Bipartite networks 152
7.3.5. Multilayer networks 153
7.4. Using social networks for recommendation 154
7.4.1. Social filtering 154
7.4.2. Extension to use attributes 155
7.4.3. Remarks 156

7.5. Experiments 156
7.5.1. Performance evaluation 156
7.5.2. Datasets 157
7.5.3. Analysis of one-mode projected networks 158
7.5.4. Models evaluated 160
7.5.5. Results 160
7.6. Perspectives 163
7.7. References 163

Chapter 8. Attributed Networks Partitioning Based on Modularity Optimization 169
David COMBE, Christine LARGERON, Baptiste JEUDY,
Françoise FOGELMAN-SOULIÉ and Jing WANG

8.1. Introduction 169
8.2. Related work 171
8.3. Inertia based modularity 172
8.4. I-Louvain 174
8.5. Incremental computation of the modularity gain 176
8.6. Evaluation of I-Louvain method 179
8.6.1. Performance of I-Louvain on artificial datasets 179
8.6.2. Run-time of I-Louvain 180
8.7. Conclusion 181
8.8. References 182

Part 4. Clustering 187

Chapter 9. A Novel Clustering Method with Automatic Weighting of Tables and Variables 189
Rodrigo C. DE ARAÚJO, Francisco DE ASSIS TENORIO DE CARVALHO
and Yves LECHEVALLIER

9.1. Introduction 189
9.2. Related Work 190
9.3. Definitions, notations and objective 191
9.3.1. Choice of distances 192
9.3.2. Criterion W measures the homogeneity of the partition P on the set of tables 193
9.3.3. Optimization of the criterion W 195
9.4. Hard clustering with automated weighting of tables and variables . . . 196
9.4.1. Clustering algorithms **MND–W** and **MND–WT** 196
9.5. Applications: UCI data sets 201
9.5.1. Application I: Iris plant 201
9.5.2. Application II: multi-features dataset 204

9.6. Conclusion 206
9.7. References 206

Chapter 10. Clustering and Generalized ANOVA for Symbolic Data Constructed from Open Data 209
Simona KORENJAK-ČERNE, Nataša KEJŽAR and Vladimir BATAGELJ

10.1. Introduction 209
10.2. Data description based on discrete (membership) distributions 210
10.3. Clustering 212
10.3.1. TIMSS – study of teaching approaches 215
10.3.2. Clustering countries based on age–sex distributions of their populations 217
10.4. Generalized ANOVA 221
10.5. Conclusion 225
10.6. References 226

List of Authors 229

Index 233