

Contents

1	On Some Facets of the Partition Set of a Finite Set	1
1.1	Lattice of Partition Set of a Finite Set.	1
1.1.1	Definition and General Properties.	1
1.1.2	Countings	9
1.2	Partitions of an Integer	18
1.2.1	Generalities	18
1.2.2	Representations	20
1.3	Type of a Partition and Cardinality of the Associated Equivalence Binary Relation	21
1.4	Ultrametric Spaces and Partition Chain Representation	30
1.4.1	Definition and Properties of Ultrametric Spaces	30
1.4.2	Partition Lattice Chains of a Finite Set and the Associated Ultrametric Spaces	33
1.4.3	Partition Lattice Chains and the Associated Ultrametric Preordonances.	37
1.4.4	Partition Hierarchies and Dendrograms	39
1.4.5	From a <i>Symmetrical</i> Binary Hierarchy to a <i>Directed</i> Binary Hierarchy	45
1.5	Polyhedral Representation of the Partition Set of a Finite Set	52
	References	58
2	Two Methods of Non-hierarchical Clustering.	61
2.1	Preamble.	61
2.2	Central Partition Method.	62
2.2.1	Data Structure and Clustering Criterion.	62
2.2.2	Transfer Algorithm and Central Partition.	69
2.2.3	Objects with the Same Representation.	72
2.2.4	Statistical Asymptotic Analysis	74
2.2.5	Remarks on the Application of the Central Partition Method and Developments	78

- 2.3 Dynamic and Adaptive Clustering Method 80
 - 2.3.1 Data Structure and Clustering Criterion 80
 - 2.3.2 The K -Means Algorithm 84
 - 2.3.3 Dynamic Cluster Algorithm 86
 - 2.3.4 Following the Definition of the Algorithm 91
- References 98
- 3 Structure and Mathematical Representation of Data 101**
 - 3.1 Objects, Categories and Attributes 101
 - 3.2 Representation of the Attributes of Type I 103
 - 3.2.1 The Boolean Attribute 104
 - 3.2.2 The Numerical Attribute 105
 - 3.2.3 Defining a Categorical Attribute from a Numerical One 107
 - 3.3 Representation of the Attributes of Type II 109
 - 3.3.1 The Nominal Categorical Attribute 110
 - 3.3.2 The Ordinal Categorical Attribute 113
 - 3.3.3 The Ranking Attribute 116
 - 3.3.4 The Categorical Attribute Valuated by a Numerical Similarity 118
 - 3.3.5 The Valuated Binary Relation Attribute 120
 - 3.4 Representation of the Attributes of Type III 121
 - 3.4.1 The Preordonance Categorical Attribute 121
 - 3.4.2 The Taxonomic Categorical Attribute 124
 - 3.4.3 The Taxonomic Preordonance Attribute 129
 - 3.4.4 Coding the Different Attributes in Terms of Preordonance or Similarity Categorical Attributes 132
 - 3.5 Attribute Representations When Describing a Set \mathcal{C} of Categories 137
 - 3.5.1 Introduction 137
 - 3.5.2 Attributes of Type I 138
 - 3.5.3 Nominal or Ordinal Categorical Attributes 138
 - 3.5.4 Ordinal (preordonance) or Numerical Similarity Categorical Attributes 142
 - 3.5.5 The Data Table: A Tarski System \mathcal{T} or a Statistical System \mathcal{S} 143
- References 146
- 4 Ordinal and Metrical Analysis of the *Resemblance Notion* 149**
 - 4.1 Introduction 149
 - 4.2 Formal Analysis in the Case of a Description of an Object Set \mathcal{O} by Attributes of Type I; Extensions 152
 - 4.2.1 Similarity Index in the Case of Boolean Data 152
 - 4.2.2 Preordonance Associated with a Similarity Index in the Case of Boolean Data 165

- 4.3 Extension of the Indices Defined in the Boolean Case to Attributes of Type II or III 178
 - 4.3.1 Introduction 178
 - 4.3.2 Comparing Nominal Categorical Attributes 180
 - 4.3.3 Comparing Ordinal Categorical Attributes 183
 - 4.3.4 Comparing Preordnance Categorical Attributes 191
- References 196
- 5 Comparing Attributes by *Probabilistic and Statistical Association I*** 199
 - 5.1 Introduction 199
 - 5.2 Comparing Attributes of Type I for an Object Set Description by the *Likelihood Linkage Analysis* Approach 201
 - 5.2.1 The Boolean Case 201
 - 5.2.2 Comparing Numerical Attributes in the *LLA* approach 221
 - 5.3 Comparing Attributes for a Description of a Set of Categories 233
 - 5.3.1 Introduction 233
 - 5.3.2 Case of a Description by Boolean Attributes 234
 - 5.3.3 Comparing Distributions of Numerical, Ordinal Categorical and Nominal Categorical Attributes 242
 - References 247
- 6 Comparing Attributes by a *Probabilistic and Statistical Association II*** 251
 - 6.1 Introduction 251
 - 6.2 Comparing Attributes of Type II for an Object Set Description; the *LLA* Approach 252
 - 6.2.1 Introduction; Alternatives in Normalizing Association Coefficients 252
 - 6.2.2 Comparing Two Ranking Attributes 256
 - 6.2.3 Comparing Two Nominal Categorical Attributes 261
 - 6.2.4 Comparing Two Ordinal Categorical Attributes 276
 - 6.2.5 Comparing Two Valuated Binary Relation Attributes 286
 - 6.2.6 From the Total Association to the Partial One 309
 - References 321
- 7 Comparing Objects or Categories Described by Attributes** 325
 - 7.1 Preamble 325
 - 7.2 Comparing Objects or Categories by the *LLA* Method 328
 - 7.2.1 The Outline of the *LLA* Method for Comparing Objects or Categories 328
 - 7.2.2 Similarity Index Between Objects Described by Numerical or Boolean Attributes 331

7.2.3 Similarity Index Between Objects Described by Nominal or Ordinal Categorical Attributes 334

7.2.4 Similarity Index Between Objects Described by Preordnance or Valuated Categorical Attributes 338

7.2.5 Similarity Index Between Objects Described by Taxonomic Attributes. A Solution for the Classification Consensus Problem 341

7.2.6 Similarity Index Between Objects Described by a Mixed Attribute Types: Heterogenous Description 344

7.2.7 The Goodall Similarity Index. 345

7.2.8 Similarity Index Between Rows of a Juxtaposition of Contingency Tables 349

7.2.9 Other Similarity Indices on the Row Set \mathbb{I} of a Contingency Table. 353

References 355

8 The Notion of “Natural” Class, Tools for Its Interpretation.

The Classifiability Concept 357

8.1 Introduction; Monothetic Class and Polythetic Class 357

8.1.1 The Intuitive Approaches of Beckner and Adanson; from Beckner to Adanson 360

8.2 Discriminating a Cluster of Objects by a Descriptive Attribute 363

8.2.1 Introduction 363

8.2.2 Case of Attributes of Type I: Numerical and Boolean 364

8.2.3 Discrimination a Partition by a Categorical Attribute 366

8.3 “Responsibility” Degree of an Object in an Attribute Cluster Formation 369

8.3.1 \mathcal{A} is Composed of Attributes of Type I. 370

8.3.2 The Attribute Set \mathcal{A} is Composed of Categorical or Ranking Attributes 375

8.4 Rows or Columns of Contingency Tables 377

8.4.1 Case of a Single Contingency Table 377

8.4.2 Case of an Horizontal Juxtaposition of Contingency Tables. 381

8.5 On Two Ways of Measuring the “Importance” of a Descriptive Attribute 382

8.5.1 Introduction 382

8.5.2 Comparing Clustering “Importance” and Projective “Importance” of a Descriptive Attribute. 386

- 8.6 Crossing Fuzzy Categorical Attributes or Fuzzy Classifications (Clusterings) 391
 - 8.6.1 General Introduction 391
 - 8.6.2 Crossing Net Classifications; Introduction to Other Crossings 394
 - 8.6.3 Crossing a Net and a Fuzzy Dichotomous Classifications 400
 - 8.6.4 Crossing Two Fuzzy Dichotomous Classifications 404
 - 8.6.5 Crossing Two Typologies 408
 - 8.6.6 Extension to Crossing Fuzzy Relational Categorical Attributes 411
- 8.7 Classifiability 419
 - 8.7.1 Introduction 419
 - 8.7.2 Discrepancy Between the Preordonance Structure and that Ultrametric, on a Data Set. 420
 - 8.7.3 Classifiability Distribution Under a Random Hypothesis of Non-ultrametricity 426
 - 8.7.4 The Murthag Contribution 431
- References 432
- 9 Quality Measures in Clustering. 435**
 - 9.1 Introduction 435
 - 9.2 The Direct Clustering Approach: An Example of a Criterion. 438
 - 9.2.1 General Presentation 438
 - 9.2.2 An Example 440
 - 9.3 Quality of a Partition Based on the Pairwise Similarities 443
 - 9.3.1 Criteria Based on a Data Preordonance 444
 - 9.3.2 Approximating a Symmetrical Binary Relation by an Equivalence Relation: The Zahn Problem 451
 - 9.3.3 Comparing Two Basic Criteria. 456
 - 9.3.4 Distribution of the Intersection Criterion on the Partition Set with a Fixed Type 468
 - 9.3.5 Extensions of the Previous Criterion 474
 - 9.3.6 “Significant Levels” and “Significant Nodes” of a Classification Tree 483
 - 9.4 Measuring the Fitting Quality of a Partition Chain (Classification Tree) 489
 - 9.4.1 Introduction 489
 - 9.4.2 Generalization of the Set Theoretic and Metrical Criteria 490
 - 9.4.3 Distribution of the Cardinality of the Graph Intersection Criterion 493

- 9.4.4 Pure Ordinal Criteria: The Lateral Order
and the Lexicographic Order Criteria 502
- 9.4.5 Lexicographic Ranking and Inversion Number
Criteria 504
- References 511
- 10 Building a Classification Tree 513**
- 10.1 Introduction 513
- 10.2 “Lexicographic” Ordinal Algorithm 519
 - 10.2.1 Definition of an Ultrametric Preordonance Associated
with a Preordonance Data 519
 - 10.2.2 Algorithm for Determining ω_u Defined by the H
Function 521
 - 10.2.3 Property of Optimality 523
 - 10.2.4 Case Where ω Is a Total Ordonance. 524
- 10.3 Ascendant Agglomerative Hierarchical Clustering Algorithm;
Classical Aggregation Criteria 527
 - 10.3.1 Preamble. 527
 - 10.3.2 “Single Linkage”, “Complete Linkage”
and “Average Linkage” Criteria 528
 - 10.3.3 “Inertia Variation (or Ward) Criterion” 530
 - 10.3.4 From “Lexicographic” Ordinal Algorithm
to “Single Linkage” or “Maximal Link”
Algorithm 534
- 10.4 *AAHC* Algorithms; Likelihood Linkage Criteria 535
 - 10.4.1 Family of Criteria of the Maximal Likelihood
Linkage. 535
 - 10.4.2 Minimal Likelihood Linkage and Average Likelihood
Linkage in the *LLA* Analysis 545
- 10.5 *AAHC* for Clustering Rows or Columns of a Contingency
Table 549
 - 10.5.1 Introduction. 549
 - 10.5.2 Chi Square Criterion: A Transposition of the Ward
Criterion 550
 - 10.5.3 Mutual Information Criterion 552
- 10.6 Efficient Algorithms in Ascendant Agglomerative Hierarchical
Classification (Clustering) 555
 - 10.6.1 Introduction. 555
 - 10.6.2 Complexity Considerations of the Basic *AAHC*
Algorithm 558
 - 10.6.3 Reactualization Formulas in the Cases of Binary
and Multiple Aggregations 560
 - 10.6.4 Reducibility, Monotonic Criterion, Reducible
Neighborhoods and Reciprocal Nearest
Neighborhoods. 566

10.6.5	Ascendant Agglomerative Hierarchical Clustering (AAHC) Under a Contiguity Constraint	572
10.6.6	Ascendant Agglomerative Parallel Hierarchical Clustering	576
	References	580
11	Applying the LLA Method to Real Data	583
11.1	Introduction: the CHAVL Software (Classification Hiérarchique par Analyse de la Vraisemblance des Liens)	583
11.2	Real Data: Outline Presentation of Some Processings	586
11.3	Types of Child Characters Through Children's Literature. . . .	590
11.3.1	Preamble: Technical Data Sheet	590
11.3.2	General Objective and Data Description	591
11.3.3	Profiles Extracted from the Classification Tree on \mathcal{A}	593
11.3.4	Developments	596
11.3.5	Standardized Association Coefficient with Respect to the Hypergeometric Model.	596
11.3.6	Return to Individuals	597
11.4	Dayhoff, Henikoffs and LLA Matrices for Comparing Proteic Sequences	600
11.4.1	Preamble: Technical Data Sheet	600
11.4.2	Introduction.	601
11.4.3	Construction of the Dayhoff Matrix	603
11.4.4	The Henikoffs Matrix: Comparison with the Dayhoff Matrix	612
11.4.5	The LLA Matrices	616
11.4.6	LLA Similarity Index on a Set of Proteic Aligned Sequences	619
11.4.7	Some Results.	625
11.5	Specific Results in Clustering Categorical Attributes by LLA Methodology	629
11.5.1	Structuring the Sets of Values of Categorical Attributes	629
11.5.2	From Total Associations Between Categorical Attributes to Partial Ones	632
	References	637
12	Conclusion and Thoughts for Future Works	639
12.1	Contribution to Challenges in Cluster Analysis	639
12.2	Around Two Books Concerning <i>Relational</i> Aspects	641
12.3	Developments in the Framework of the LLA Approach	643
12.3.1	Principal Component Analysis	643
12.3.2	Multidimensional Scaling	644

12.3.3	In What <i>LLA</i> Hierarchical Clustering Method Is a Probabilistic Method?	645
12.3.4	Semi-supervised Hierarchical Classification	645
12.4	Big Data	646
	References	646



<http://www.springer.com/978-1-4471-6791-4>

Foundations and Methods in Combinatorial and
Statistical Data Analysis and Clustering

Lerman, I.C.

2016, XXIV, 647 p. 54 illus., Hardcover

ISBN: 978-1-4471-6791-4