

Jan W. Owsinski

Data Analysis in Bi-partial Perspective: Clustering and Beyond

Contents

1	Notation and Main Assumptions	1
1.1	Notation	1
1.2	Situation Studied and Its Characterisation	3
2	The Problem of Cluster Analysis	9
2.1	The General Formulation	9
2.2	The Meaning of the Problem	10
2.3	The Issues of Practicality	11
2.3.1	The Nature of Partition	11
2.3.2	Similarity Inside and Dissimilarity Outside: How Do the Existing Methods Fare?	13
2.3.3	Defining a Cluster	17
2.3.4	The Use of Metaheuristics	18
2.3.5	The Number of Clusters	18
2.3.6	The Shapes of Clusters	19
2.3.7	Why not the Exhaustive Search?	19
2.3.8	On Clustering Validity Indices	20
	References	21
3	The General Formulation of the Objective Function	23
3.1	The Formulation	23
3.2	Some Explanations and the Leading Example	24
3.3	A Broader View: The Levels of Perception and the Issue of Scale	29
4	Formulations and Rationales for Other Problems in Data Analysis	33
4.1	Categorisation	33
4.2	Optimum Histogram	36
4.3	Division of the Univariate Distribution	38
4.4	The p-Median/p-Center Facility Location Problem	47

4.5	Block-Diagonalisation or Concept Identification	51
4.6	Rule Extraction	54
4.7	A More General Analogy	58
4.8	Minimum Message Length	60
4.9	Number of Factors in Factor Analysis	63
4.10	Ordering—Preference Aggregation	64
	References	66
5	Formulations in Cluster Analysis	69
5.1	The Leading Example Again	69
5.1.1	The Formulation Repeated	69
5.1.2	The Reference to the Relational Form of the Bi-partial Objective Function and the Respective Solution	71
5.1.3	The Prerequisites for the Algorithm Relative to the MP Formulation	72
5.2	The Bi-partial Version of the k-means Algorithm	75
5.2.1	The Standard Case	75
5.2.2	Two Illustrative Examples	77
5.3	Some Other Implementations of the Bi-partial Objective Function	79
5.3.1	Preliminary Remarks—The Objective Function and the Algorithm	79
5.3.2	The Additive Objective Function with a Constant Cluster Cost	80
5.3.3	The Case of Minimum Distances and Maximum Proximities	81
5.3.4	The Case of Average Distances and Additive Proximities	82
5.4	Comparing and Assessing Clustering Quality: The Indices and the Principles	82
5.4.1	Introductory Remarks	82
5.4.2	The Exemplary Internal Clustering Quality Measures	84
5.4.3	The Founding Ideas	87
5.4.4	Assessing Clustering Quality	90
5.5	Summarising the Place of the Bi-partial Approach and the Algorithms Thereof	92
	References	93
6	The General Sub-optimisation Algorithm and Its Implementations	97
6.1	Basic Properties of the Objective Function	97
6.2	The General Sub-optimisation Procedure	100
6.2.1	The Algorithmic Form of the Objective Function	100
6.2.2	The General Form of the Algorithm	104

6.3	Some Comments on the Procedure	105
6.3.1	Divergence from Optimality	105
6.3.2	Analogous Transformation of the Relational Form of the Objective Function	106
6.3.3	The Complexity of the Procedure	107
6.3.4	The Character of the Results and the Values of Parameter r	107
6.3.5	Possibility of Weakening of the Assumptions	108
6.3.6	Properties of the Method	108
6.3.7	The Values of the Objective Function	110
6.4	Concrete Realisations of the General Algorithm	111
6.4.1	Introductory Remarks	111
6.4.2	The Algorithm with Additive Objective Function	112
6.4.3	The Algorithm for the Objective Function with Additive Proximities and Constant Cluster Setup Cost	113
6.4.4	The Algorithm of the Extreme Distances and Proximities	114
6.5	The Properties of the Algorithmic Realisations	116
6.5.1	Aggregation According to Minimum Distance	116
6.5.2	Objective Function Additive with Respect to Clusters	117
6.5.3	Additivity Along Hierarchy	118
6.5.4	Distances and Proximities with Respect to Centres	119
6.6	Algorithms for the Objective Functions Additive with Respect to Clusters	119
6.6.1	Algorithm of Mean Distances and Additive Proximities	119
6.6.2	Algorithm of Minimal Distances and Additive Proximities	120
6.6.3	Algorithm of Maximal Distances and Additive Proximities	120
6.6.4	Algorithm of Partition Cardinality and Additive Proximities	121
6.6.5	Algorithm of Additive Distance and Minimum Cardinality	121
6.6.6	Algorithm of Additive Distance and Sum of Means	122
6.7	Algorithms of the Objective Function Non-additive with Respect to Clusters	122
6.7.1	Introduction	122
6.7.2	Algorithm of the Range of Inter-cluster Distance Values	123
6.7.3	Algorithm of the Sum of Distances from the Centre of Partition	123
6.7.4	Algorithm of Variance of the Inter-cluster Distances	124

6.8	Designing an Algorithm	125
	References	131
7	Application to Preference Aggregation	133
7.1	Introductory Remarks and the Setting	133
7.2	The Linear Programming Formulation	134
	7.2.1 The Linear Program	134
	7.2.2 The Parameterisation	135
7.3	The Alternative Algorithmic Approach.	137
7.4	An Illustration	139
7.5	Some Conclusions	141
	References	143
8	Final Remarks	145
	Bibliography	147
	Index	151