Studies in Computational Intelligence

Volume 818

Series Editor
Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
The series “Studies in Computational Intelligence” (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

The books of this series are submitted to indexing to Web of Science, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink.

More information about this series at http://www.springer.com/series/7092
Jan W. Owsiński

Data Analysis in Bi-partial Perspective: Clustering and Beyond

Springer
The considerations forwarded in the present volume suggest that the use of the general “bi-partial” objective function leads to a more effective capacity of solving the problems of cluster analysis in that it implies obtaining both the cluster content and the cluster number, without recurring to an “external” criterion. The method, proposed, embraces also the possibility of designing a relatively simple and intuitive algorithm of sub-optimisation. The same principles can be similarly successfully applied to several other essential problems of data analysis.

As a “by-product” of the developments, related to the paradigm of the bi-partial objective function, several important issues surface in a natural manner and get an interesting interpretation and/or explanation. One of those is the matter of “scale”, crucial for the understanding of the adequacy of partitions in clustering, here intimately associated with the transformation between distance and proximity. Within the bi-partial approach, this “scale” can be made explicit and subject to choice.

Another such “by-product” is constituted by the association with the progressive merger (agglomerative) techniques, classical to clustering, the respective merger rules, minimum distance principle, etc. Here, definitely, the bi-partial approach provides a much deeper insight into the principles and structures involved.

Regarding clustering, the book considers solely the clustering problem proper, so that a lot of related research issues are left untreated, like, the problem of mixture of distributions or the specific problems associated with pattern recognition. On the other hand, the content of this book does not address at all the currently dominating meta-heuristic approaches, applied in clustering, since they refer uniquely to the algorithmic side of finding the solutions to the clustering problem, and all of them have, anyway, to be guided by the more general principles as to what such solutions actually are. And this is exactly what the study here reported deals with.

It should be noted that the book is “technical” in the sense that it occupies a place between computer science type of considerations and those belonging to applied mathematics, emphasis being on practical and effective solving of problems having definite substance matter contents, whose interpretation is of primary importance.
An important reservation ought to be made at this point: \textit{the book is not about yet another objective function for clustering}. It is about an integral approach, in which the proposed form of the general objective function, which can be instantiated for the various concrete perceptions and interpretations of the clustering problem, plays the central role.

Few words are perhaps due, concerning the relation between the content of this book and the so-called soft approaches. The adjective “soft” appears, namely, to be reserved to a definite set of methodologies, so that, for example, in clustering, the classical k-means algorithm is treated as “hard”, when compared to fuzzy c-means (FCM), qualified as “soft”. This is, of course, a false image, since clustering as such should be in many instances considered to be a “soft” approach (unless it is not only treated, but actually applied as just a preliminary stage) to such problems as model identification, facility location, cell formation in flexible manufacturing, information retrieval and many others. For virtually all of these, mathematical programming formulations exist or can relatively easily be developed, possibly accompanied by sensitivity analysis. Hence, a difference between these two ways of approaching the respective problems can easily be established. Even within the data analysis domain, most of clustering approaches (including naturally those based on fuzzy or intuitionistic precepts) should be considered “soft” in the sense that the respective problem and the way to solve it are not very precisely defined, with a wide margin for interpretation of both problem statement and the results obtained.

The content of this report has the following structure: first, notations used in the study are introduced and their context (“problem background”) is explained in Chap. 1. Then, in Chap. 2, the generic problem of cluster analysis is formulated, along with all the consequences thereof, especially the indications of necessary complements, needed to make the problem of clustering practically tractable.

In Chap. 3, the general formulation of the objective function is introduced and justified, along with some basic illustrations and examples. The rationale for and the ways of constructing this objective function are amply illustrated with examples of its concrete implementations, satisfying the prerequisites previously proposed, in Chap. 4 for a number of diverse standard problems in data analysis, but also bordering upon other domains, e.g. operational research. These examples are then complemented in Chap. 5 by some further instances, belonging, however, uniquely to the domain of cluster analysis. This chapter also contains the consideration of other criteria and evaluation functions, used in various problems of data analysis, which are related to the here introduced general two-sided objective function and to derivation of the connections between them. The chapter ends with some remarks and suggestions on the way the clustering methods and results ought to be assessed.

The algorithm, which sub-optimises with respect to the general objective function, is introduced, against the background of the required properties, along with the concrete versions of this algorithm for a spectrum of the exemplary implementations of the objective function, in Chap. 6. It is also shown there that the algorithms obtained for the bi-partial objective function share some of the essential characteristics with the classical hierarchical progressive merger algorithms, described with the well-known Lance–Williams formula. Being in a way similar to
those algorithms, the ones proper for the bi-partial objective function provide for the stop condition, concerning the merger iterations, as well as a natural index of hierarchy.

In Chap. 7, the application is shown of the entire approach and its philosophy, including also the possibility of deriving an effective algorithm, to quite a specific, and partly different case, namely the one of preference (precedence) aggregation. Following the final remarks, forming Chap. 8, and a concise index, the volume ends with a broad bibliography, related, on the one hand, primarily to cluster analysis and the works using or based on explicit formulations of objective functions, and on the other hand—to those other issues in data analysis, where similar principles are observed or at least postulated. A separate list of references is put together, concerning the work of the present author, essentially concerning the bi-partial approach, so as to not make an impression of dominating the references altogether.

The book is meant for data science specialists, who would like to broaden their perspective on the fundamental approaches available, but first of all—on the way many problems are perceived in and thus to find answers to some questions usually either overlooked or solved via cumbersome or not fully convincing manners. It is also meant for graduate students, dealing with computer and data sciences, who might wish to complement their knowledge and skills with a fresh insight into many problems, otherwise treated in some standard “academic” manners.

The Readers, who might be interested in the “historical” roots of the approach here presented, are encouraged to take, perhaps, first a look at the Sect. 5.4.3 of the volume, especially if they are already somewhat knowledgeable in the domain of clustering.

Warsaw, Poland

Jan W. Owsiński
The present book results from the studies, conducted by the author, concerning data analysis, and especially cluster analysis and preference aggregation. The volume contains the general formulation, the properties, the examples and the techniques associated with a general objective function, referred to as “bi-partial” objective function, devised mainly for purposes of effective solving of the clustering problems.

This objective function is based on the principle of simultaneous consideration of two aspects of clustering: the intra-cluster similarity and inter-cluster dissimilarity (or, dually, intra-cluster distance and inter-cluster proximity). This sounds, definitely, very much obviously, and even perhaps bordering upon triviality. Yet, it is shown here how the general principles of construction of such an objective function can be implemented through concrete, practically applicable formulations. Further, it is demonstrated that both the general form of the function and its concrete implementations imply the solutions to the clustering problem in terms of both the number of clusters and their composition. Then, a general algorithm is proposed, leading to the sub-optimal solutions with respect to the objective function, through a special type of classical progressive merger procedure. The properties of this algorithm and the examples of its concrete implementations are also presented.

The key property that is stressed in the methodology here proposed is the adequate representation of the generic problem of cluster analysis. It is shown that most, if not all of the existing approaches fail with this respect, and that is why they usually also fail to propose, within the same approach, a consistent and effective algorithm of finding the solutions to the clustering problem.

It is also shown how the general objective function can be formulated in a relatively easy, but also constructive and effective manner for quite a wide variety of problems in multivariate data analysis (e.g. categorisation, optimum histogram and rule extraction). Likewise, examples are shown of its equivalence to some kinds of quality criteria or indices used also, or at least referred to in various kinds of data
analytic problems. In particular, application of the fundamental principles of con-
struction of the objective function and of the algorithm to the problem of preference
aggregation is presented.

Warsaw, Poland

Jan W. Owsiński
## 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
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 .......................... 109
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
List of Figures

Fig. 2.1 An example: furniture objects described by two variables 10
Fig. 2.2 An example of a set of objects, for which overlapping clusters may be justified 12
Fig. 2.3 An example of a set of objects featuring complex clusters shapes 20
Fig. 3.1 A simple example of a data set 27
Fig. 3.2 Values of the bi-partial objective function for the data of Fig. 3.1 and the sequence of partitions given in the text 27
Fig. 3.3 Another simple example of a data set, similar to that of Fig. 3.1 28
Fig. 3.4 Values of the bi-partial objective function for the data of Fig. 3.3 and the sequence of partitions given in the text 28
Fig. 3.5 Yet another—similar—simple example of a data set 28
Fig. 3.6 The values of the bi-partial objective function for the data of Fig. 3.5 and specifications as in the respective text 29
Fig. 3.7 An illustration for the significance of the perception of scale 32
Fig. 4.1 An illustration of data for the task of categorisation 34
Fig. 4.2 Complete histogram of variable values from Fig. 4.1 36
Fig. 4.3 Histogram for the case of Fig. 4.1 with three intervals 37
Fig. 4.4 Histogram for the case of Fig. 4.1 with four intervals 37
Fig. 4.5 Another four-segment-based histogram for the case of Fig. 4.1 37
Fig. 4.6 An academic example of a convex cumulative distribution 39
Fig. 4.7 The scheme of determination of similarities between the two consecutive approximating segments 43
Fig. 4.8 An example of the total quality of life scoring, see http://www.il-ireland.com/il/qofl07/ 45
Fig. 4.9 Results (time in seconds) of the runners in the marathon race (“Koral marathon”) in Krynica, Poland, in 2012 (diagram shows the data published in “Rzeczpospolita” daily) 46
Fig. 4.10 A simple example for the facility location problem ............. 49
Fig. 4.11 A simple example of an incidence matrix ....................... 51
Fig. 4.12 Incidence matrix from Fig. 4.10 after a permutation of rows and columns aiming at block-diagonalisation .................... 51
Fig. 4.13 Another view at the blocks from Fig. 4.12 ...................... 53
Fig. 4.14 Various evaluations of partitions into blocks for the example of Figs. 4.11, 4.12 and 4.13 ........................................... 54
Fig. 4.15 A scheme of analogous interpretations related to basic criteria in classification, block-diagonalisation and rule extraction; a, b, c and d denote numbers of respective objects ................... 58
Fig. 4.16 An illustration of various situations arising in connection with indiscernibility applied to MML/MDL type problem of clustering ................................................................. 62
Fig. 4.17 Schematic view of the typical distribution of membership values in a (fuzzy) cluster q, obtained, e.g., with an FCM algorithm .......................................................... 63
Fig. 4.18 Schematic view of the basis for the bi-partial setting to solve the problem of the number of factors in factor analysis .......... 65
Fig. 5.1 Academic Example 1 of a data set for the bi-partial k-means ................................................................. 77
Fig. 5.2 Academic Example 2 of a data set for the bi-partial k-means ................................................................. 78
Fig. 5.3 An example regarding Regnier’s idea for n = 2, i.e. the unidimensional case; the value of proximity s12 indicates that joining of the two objects is more advisable than separating them (abstraction is made here of the normalisation issue) .... 88
Fig. 5.4 An example regarding Regnier’s idea for n = 3; intuitively, the proper solution appears to be {{1, 2}, 3} ......................... 88
Fig. 5.5 An example of data for case (1) of the general references ....... 91
Fig. 5.6 An example of data for case (2) of the general references ..... 92
Fig. 5.7 Schematic view of the place of bi-partial approach and the potential effective algorithms against the background of the domain of data analysis .................... 93
Fig. 6.1 Schematic view of the proposed “opposite monotonicity” of the global components of the bi-partial objective function. The points indicate the maxima over all partitions for consecutive values of p ........................................ 99
Fig. 6.2 Illustration for the convexity of the function Q_{D}^{S}(r)Q_{D}^{P}(P^{**}) - Q_{S}(P^{**}) > Q_{S}(P^{*}) - Q_{D}^{P}(P^{*}) > Q_{D}^{P}(P^{**}) - Q_{S}(P^{***}) : .......................................... 102
Fig. 6.3 An illustration for the outlook and formation of the function Q_{D}^{S}(r) ......................................................... 103
<table>
<thead>
<tr>
<th>Fig.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4</td>
<td>A simplistic case of a cube, with 27 objects, positioned equidistantly on the surfaces and in the centre of the cube. The distances between the neighbouring objects are equal 1</td>
<td>128</td>
</tr>
<tr>
<td>6.5</td>
<td>Ten objects, situated on a single line, with, again distances between neighbours equal 1</td>
<td>129</td>
</tr>
<tr>
<td>6.6</td>
<td>Twelve objects, situated on a single line, with distances between the closest neighbours equal 1</td>
<td>129</td>
</tr>
<tr>
<td>6.7</td>
<td>Another simple example, again with distance between neighbours equal 1</td>
<td>129</td>
</tr>
<tr>
<td>6.8</td>
<td>Distribution of distances for the example from Fig. 6.4</td>
<td>129</td>
</tr>
<tr>
<td>6.9</td>
<td>Distribution of distances for the example from Fig. 6.5</td>
<td>130</td>
</tr>
<tr>
<td>6.10</td>
<td>Distribution of distances for the example from Fig. 6.6</td>
<td>130</td>
</tr>
<tr>
<td>6.11</td>
<td>Distribution of distances for the example from Fig. 6.7</td>
<td>130</td>
</tr>
</tbody>
</table>
List of Tables

Table 4.1 Examples of division of the cumulative distribution from Fig. 4.6 and the corresponding values of $C_D\{A_q\}$. .............. 41
Table 4.2 Examples of division of the cumulative distribution from Fig. 4.6 and Table 4.1, and the corresponding values of $C_D\{A_q\}$, $C_S\{A_q\}$, and $C_{DS}\{A_q\}$. .................. 44
Table 4.3 Values of $Q_{DS}^2(P) = Q_D(P) + Q_S(P)$ for a series of partitions, according to (4.17) .................................. 50
Table 4.4 An example for the rule extraction problem ................. 57
Table 4.5 An example for the rule extraction problem: separation of observations for combinations of “output variable” values .... 57
Table 5.1 Optimum values of the objective function (5.9) and its components for the successive numbers of clusters for the academic Examples 1 and 2, illustrated in Figs. 5.1 and 5.2 .................................................................. 78
Table 6.1 Selected examples of algorithms for the instances of formulations of $Q_D$ and $Q_S$ additive with respect to clusters .............................................................. 118
Table 6.2 Illustration for the approximate fulfilment of condition (6.69). ................................................................. 126
Table 6.3 The bounds on the cardinality of the dominating cluster, $n_1^\text{min}(p)$ and $n_1^\text{max}(p)$, for increasing values of $p$ in the case of $n = 100$, ensuring fulfilment of (6.69) with the sign “≥”. .... 127
Table 7.1 An instance of the set of match scores for a small soccer cup. ............................................ 139
Table 7.2 Comparison of three orderings: resulting from the cup “rules of the game”, match points and goal scores .................. 141
Table 7.3 Match results shown for the ordering according to the “rules of the game” ........................................ 141