Control
and Cybernetics
VOL. 15 (1986) No. 2

Optimization in clustering: an approach and other approaches

by

JAN W. OWSIŃSKI

Systems Research Institute,
Polish Academy of Sciences
Newelska 6, 01-447 Warszawa

The paper provides an introduction to the special issue of "Control and Cybernetics", devoted to optimization formulations and approaches in cluster analysis, by outlining an approach to clustering based upon a simple formulation of objective function yielding globally optimal partitions, and then presenting a short overview of some of the other problem formulations and approaches, in particular those shown in this issue.

1. Introduction

The present short introductory paper is subdivided into two distinct parts: in the first part an approach to clustering is outlined together with its "historical" development, while in the second part some other approaches are superficially reviewed against the background of the one presented in the first part. Thus, the author takes advantage of his position as the guest editor of this issue to place in this perspective the papers contained herein.

* * *

Let us first repeat a few banalities. That putting together or associating similar and distinguishing or discriminating dissimilar is a basic human intellectual activity, necessary for scientific inquiry and, far before, for language creation, seems quite obvious, Hempel (1952), Kruskal (1977) or even Linnaeus (1737). That the process of associating and distinguishing is successfully performed by the human race is not only witnessed by effective use of language (notwithstanding its abuses, anyway corrected over longer time periods), but also by the experience in practical activities, directly and indirectly based upon associating and distinguishing. In fact, this process leads to creation, however vaguely defined they may be, of models of reality, at first very rough, but also very robust ones, and then more refined ones. The

need for the latter ones arises when the first ones had already been created and used and their effectiveness is no longer satisfactory. Establishment, however, of the more refined models is usually not possible with the same sort of simple intellectual process — for a more refined model a more refined approach is needed. There is, though, it seems, a law of decreasing returns to these efforts, as well. Thus, not only a refined model must be built upon a simple one, but also it may require more input to get a comparable result.

Now, associating and distinguishing is all what cluster analysis is about.

In accordance with the previous remarks cluster analysis is the first thing to do when confronted with a data set "to analyse". In order, however, for cluster analysis to be the domain and tool of scientific insight, the simple phrase "associating and distinguishing" must be more precisely worded.

2. The problem and a solution

The problem is:

given a set of objects numbered i, $i \in I=\{1, ..., n\}$, these objects being characterized in such way that it is possible to define their distances/dissimilarities and/or proximities/similarities, to find such a partition P of I, $P=\{A_q\}_{q=1}^{p}$, $\bigcup_{q=1}^{p} A_q=I$, that objects belonging to the same A_q's be possibly close/similar while those belonging to various A_q's be possibly distant/dissimilar. (1)

This is representation of "associating and distinguishing". It is sufficient for many purposes, but not yet for practical ones. Thus, cluster analysis is meant to

A. formalize (1) in its totality and in its various aspects,

B. provide solution methods to thus formalized problems.

Since, as remarked before, cluster analysis is used on a first, exploratory stage of analysis, formal statements of (1) should not foreclose the basic choices to be made regarding the potential data structure. On the other hand, wording of (1) implies a capacity of comparing goodness of various partitions. Thus, it can be deduced that *A.* should essentially provide an objective function (goodness criterion) which would imply a partition, or class of partitions, globally optimal in $E_p(I)$, i.e. in the space of all partitions of I. Naturally, there may in principle be various types of objective functions, depending upon the purpose of data exploration. Before, however, turning towards the choice of objective function, let us look at the more basic notions related to proper formulation of (1).

First, distances/proximities. Obviously, their definition, except for clarification of their basic properties, is within the responsibility of a specialist from the domain to which the data set pertain. In fact, it is not rare that a clustering method be closely related to distance/proximity definition, but in such a case the whole decision is simply moved over to the subsequent level of consideration.

This subsequent level of consideration is related to definitions of coherence or disparity for clusters (within clusters and/or among clusters). It is namely at this level that the fundamental choices are being made: is, for instance, intracluster coherence defined by the distances d_{ij} (or similarities s_{ij}) of objects i, j belonging to this cluster, $i, j, \in A_q$, by the distances d_i^q or similarities s_i^q between objects i belonging to cluster A_q and a "representation" of this cluster, and how this "representation" is defined, or, eventually, by some other cluster-proper measure involving objects' characteristics x_i, $x_i \in E_X$.

The third level, i.e. the one of aggregation of cluster-proper measures does usually entail a trivial decision, and that is why the second level is crucial for the objective function definition.

Depending on whether one wishes to "simply" group the objects, to perform canonical analysis etc., various measures can be defined on the cluster level, see e.g. paper by E. Diday in this issue. Notwithstanding various particular purposes of data analysis performed, the fundamental options must be, as said, kept possibly open. This applies primarily to the question of "optimal" cluster number (see e.g. papers by G. Libert or by Kusiak, Vannelli and Kumar in this issue). The problem is that most of the objective functions, as pointed out by G. Libert, have optimum values forming monotonic sequence in p. Namely, let $Q(P): E_P \to R$ be a clustering objective function and $Q^{\text{opt}} = Q(P^{\text{opt}})$, its optimum value attained for a globally optimum partition P^{opt}, this partition possibly understood in a broad sense, i.e. eventually also together with cluster "representation". Further, let $Q^{\text{opt}}(p) = \underset{P(p) \in \{P \mid \text{card } P = p\}}{\text{opt}} Q(P(p))$ be the optimum value of Q for partitions P with card $P = p$. Then, there is usually either $Q^{\text{opt}}(p) \geqslant Q^{\text{opt}}(p+1)$, or $Q^{\text{opt}}(p) \leqslant Q^{\text{opt}}(p+1)$, $\forall p \in \{1, ..., n\}$.

It is with mainly this problem in mind that the present author has worked with the objective function to be maximized

$$Q^0(P) = \bar{Q}_D(P) + \underline{Q}_s(P) \tag{2}$$

where $\bar{Q}_D(P)$ is the aggregate measure accounting for inter-cluster dissimilarities or distances, while $\underline{Q}_s(P)$ is the aggregate measure accounting for intra — cluster similarities or proximities, see e.g. Owsiński (1980, 1984).

For the "simple" grouping purpose, and

$$\bar{Q}_D(P) = \sum_{q'=1}^{p-1} \sum_{q''=q'+1}^{p} D_{q'\,q''}, \qquad \underline{Q}_s(P) = \sum_{q=1}^{p} S_q \tag{3}$$

and further

$$D_{q'\,q''} = \sum_{i \in A_{q'}} \sum_{j \in A_{q''}} d_{ij}, \; S_q = \sum_{\substack{i \in A_q \\ i<j}} \sum_{j \in A_q} s_{ij} \tag{4}$$

where $s_{ij} = d^{\max} - d_{ij}$, $d^{\max} = \max_{i, j \in I} d_{ij}$

(2) can easily be transformed into the objective function used by Marcotorchino and Michaud in their LP formulations, Marcotorchino and Michaud (1978, 1979):

$$Q^M(P) = \sum_{i,j} \hat{d}_{ij}(1-y_{ij}) + \sum_{i,j} (1-\hat{d}_{ij})\, y_{ij} \tag{5}$$

where $\hat{d}$ and $\hat{s}$ are normalized, $y_{ij}=1$ when i and j belong to the same cluster, and $y_{ij}=0$ otherwise, and (5) is maximised under constraints

$$y_{ij}=y_{ji}$$

$$y_{ij}+y_{jk}-y_{ki}\leqslant 1 \qquad i,j,k \in I \tag{6}$$

and

$$y_{ij} \in \{0,1\} \quad \text{or} \quad y_{ij} \in [0,1] \tag{7}$$

making y_{ij} form a partition. Constraint (7) in its second form allows obtaining of fuzzy partitions, see Bezdek et al., or Libert, in this issue. In practice, however, solutions obtained with $y_{ij} \in [0, 1]$ excepting special cases are not fuzzy.

Formulations analogous to (5, 6, 7) can be used, with some constraints modified, to problems other than clustering, e.g. to aggregation of preorders, and, in fact, to quite a wide variety of problems in which pairwise comparisons are possible. For a class of such problems Roubens (1982) remarked that this method "overcomes partially the (*existing till then*) gap, by suggesting a heuristic solution which, since, has not been essentially improved". This remark is still valid. Moreover, heuristics mentioned applies solely to a numerical approach to solving (5, 6, 7) for larger n. Otherwise, accurate solutions are obtained through standard software. Thus, treatment of (1) through (5, 6, 7) provides an adequate answer to A and also to B, the latter with exception of larger n.

On the other hand, starting with (2), the present author develops a very simple suboptimization procedure, similar to the classical progressive merger procedures. Thus, again, A and B are satisfied, but, this time, for a potential difference between the optimum and the suboptimization result. The approach can also be broadened to encompass other problems, like aggregation of preorders, see Owsiński and Zadrożny (1985), although not in such a straightforward manner as in the Marcotorchino — Michaud method.

There is a number of predecessors to the two formulations, as well as some quasi-equivalent problem statements and formulae, whose main shortcoming is the lack of clear and efficient solution algorithm. Within the sensu tricto clustering domain the first reference, having, however, more of theoretical than practical impact, is Regnier (1965), and then, independently, Fortier and Solomon (1966), and Rubin (1967). Ducimetière (1970) provided the first hint as to the possible, and feasible, solution procedure, but stopped short of developing it. De Falguerolles (1978) gave the analysis of the objective function itself going beyond Regnier and finding a number of parallels but at this point the question *B.* had still not been solved in an efficient way until the two methods mentioned here appeared. The

algorithms proposed, if any, reduced usually to exchange procedures, requiring an initial guess of partition, a P^0, and consisting in sequential checking of partitions differing from the previous one by the assignment of one, two, etc. objects to clusters. Such a procedure becomes very quickly infeasible and hence the solution thus obtained is largely determined by P^0. At a closer inspection, however, it is easy to see that such, more refined procedures, when applied to other, more appropiate objective functions, in methods like K-means of Mac Queen (see, e.g. Anderberg (1973) or Hartigan (1975)), like "dynamic clusters" of Diday (see e.g. Diday et al. (1979)), or like approaches analysed by Späth in e.g. Späth (1983), may turn out quite efficient, in spite of preservation of their essentially local nature. In fact, these procedures do in cases of some purposes of data analysis, for instance — establishment of local linear interrelations or quasimodels of data — prove to be the only feasible approaches to date.

The other line of thinking, leading to the two formulations given, related primarily to ordinal analysis, the voting problem etc., evokes first some classical analyses related to social and political sciences, starting with Condorcet (1785) and his voting aggregation rule, through Pareto (1927) and his postulates, down to Arrow (1963) and his axioms. Obviously, however, the first really constructive references in this area are provided by Kemeny and Snell, Kemeny (1959), Kemeny and Snell (1962), as well as by Kendall (1962), who provide related formulae for the objective function, and also by Barbut (1966). Further steps were made also by Berthelemy and Montjardet (1980, 1981), when, however, solution was already available.

It is, perhaps proper to also mention here the work of de la Vega (1967a, b), whose practical significance could also be assessed only when the formulations (2) and (5, 6, 7) were given adequate algorithm for solution.

One more remark. Although the methods mentioned are not meant to provide optimum hierarchies of partitions, they in fact can generate hierarchies, through modification of (2) into

$$Q^0(P, r) = r\bar{Q}_D(P) + (1-r)\underline{Q}_s(P) \tag{8}$$

The method developed by this author produces the hierarchy in a natural manner, together with index values of r, $r \in [0, 1]$, while the Marcotorchino — Michaud method can easily be made to produce hierarchy by parametrizing with regard to r as in (8), but put into (5), and by adding appropriate constraints.

3. Approaches in this volume

In the light of previous remarks and the variety of cluster analysis applications it is no wonder that several papers contained in this issue do discuss A, that is, objective function formulations, their interpretation and use, see e.g. papers by E. Diday, A. Kusiak et al., J. L. Chandon and F. F. Boctor or J. W. Owsiński

and S. Zadrożny. In fact, except for a remark in the paper by Chandon and Boctor and the background illustration in the paper by Owsiński and Zadrożny, the approaches outlined in section 2 are not recalled. (Note that these two papers are both related to ordinal data, to which these approaches are by no means limited.) This is caused partly by different, quite specific applications, e.g. Diday — intervariable relations, and Kusiak et al. or Garcia and Proth, the latter — note their mutual implications! — meant primarily for operations research purposes, and also partly by different "philosophical" foundations, see Bezdek et al. and Libert. With respect to the fuzzy set theoretic formulations and assumptions it should be repeated that e.g. the Marcotorchino — Michaud method does in fact allow, in general, non — $\{0, 1\}$ solutions, although they simply do not appear in the solutions, with exception, perhaps, of such obvious cases as the Ruspini's "butterfly". Thus, it is necessary to apply special fuzzy — partitions — oriented formulations in order to obtain the anticipated structure of results, these formulations being often related to e.g. K-means-like functions. An apparent possibility for extending such approaches into other "center and reallocate" procedures exist. Another potential area exists within the "percolation" type of methods (see Tremolières (1979)).

With regard to operations research the paper by Späth offers a problem formulation which is encountered in this area and is, in fact, an interesting quasi-dual to (1), i.e. the search for possibly alike, but internally diversified clusters.

Many of the clustering problems formulated as optimization tasks are being solved via heuristic approaches, referring most often to exchange algorithms. There is, however, a family of mathematical programming methods, whose origin goes back e.g. to Mulvey and Crowder (1979), and which are referred to in this volume, obviously, by Mulvey, but also by Kusiak, Vannelli and Kumar. The paper by Bezdek et al. presents also an explicit optimization approach, specialized for the class of fuzzy-partitions-bound objective functions.

In order to provide a background for the — explicity or implicitly — optimization — oriented papers, a work of Stańczak relying primarily upon graph — theoretic antecedents is presented. It is interesting to see the generalizations of a relatively simple and intuitively easily understandable notions into the far-reaching, although a bit less tangible ones.

Papers contained in this special issue of Control and Cybernetics are ordered in such a way that first the ones presenting more theoretical aspects and then the ones more application oriented appear.

The editor would like to express his thanks to all the contributing authors for their effort and for acceptance of conditions set. Separate thanks go to the reviewing team. As to some of those papers that for technical or other reasons could not get into this issue, it is intended that they shall be published in the forthcoming issues of the quarterly.

It is indeed our hope that the present issue shall constitute an important step in the development of the domain.

References*

[1] ANDERBERG M. R. Cluster Analysis for Applications. New York, Academic Press, 1973.

[2] ARROW K. Social Choice and Individual Value. New York, Wiley, 1963.

[3] BARBUT M. Note sur les ordres totaux à distance minimum d'une relation binaire. Math. et sciences humaines. (1966) 17, 47–48.

[4] BARTHÉLEMY J. P., MONTJARDET B. Ajustement et résumé de données relationnelles: les relations centrales. In: Data Analysis and Informatics. Amsterdam, North Holland, 1980.

[5] BARTHÉLEMY J. P., MONTJARDET B. The median procedure in cluster analysis and social choice theory. *Math. Social Sciences.* **1** (1981) 3.

[6] CONDORCET J. A. Essai sur l'application de l'analyse à la probalité des decisions rendues à la pluralité des voix. Paris, Ac. Roy. des Sciences, 1785.

[7] DIDAY E. et al. Optimisation en classification automatique. Le Chesnay, INRIA, 1979.

[8] DUCIMETIÉRE P. Les méthodes de la classification automatique. *Rev. Stat. Appl.* **18** (1970) 4, 5–25.

[9] DE FALGUEROLLES A. Classification automatique: un critère et des algorithmes d'échange. In: Classification automatique et perception par ordinateur. Le Chesnay, IRIA, 1978.

[10] FORTIER J. J., SOLOMON H. Clustering procedures. In: Multivariable Analysis I, P. R. Krishnaiah. New York, Academic Press, 1966.

[11] HARTIGAN J. A. Clustering Algorithms. New York, Wiley, 1975.

[12] HEMPEL Problem of Concept and Theory Formation in the Social Sciences, Language and Human Rights. University of Pennsylvania Press, 1952.

[13] KEMENY J. Mathematics without numbers. *Daedalus*, **88** (1959), 571–591.

[14] KEMENY J., SNELL L. Mathematical Models in the Social Sciences. Cambridge, MIT Press, 1962.

[15] KENDALL M. G. Rank Correlation Methods. London, Griffin, 1962.

[16] KRUSKAL J. The relationship between multidimensional scaling and clustering. In: Classification and Clustering, J. Van Ryzin. New York, Academic Press, 1977.

[17] LINNAEUS C. Genera Plantarum, 1737, after: Everitt B. Cluster Analysis, New York, Halsted Press, 1980.

[18] MARCOTORCHINO F., MICHAUD P. Optimization in ordinal data analysis. IBM France Scientific Centre. Technical report 001. Paris, 1978.

[19] MARCOTORCHINO F., MICHAUD P. Optimisation en analyse ordinale des données. Paris, Masson, 1979.

[20] MULVEY J. M., CROWDER H. P. Cluster analysis: An application of Lagrangean relaxation. *Mgmt. Science*, **25**, (1979), 329–340.

[21] OWSIŃSKI J. Regionalisation Revisited: and Explicit Optimization Approach: IIASA, CP-80-26, Laxenburg, 1980.

[22] OWSIŃSKI J. On a quasi-objective global clustering method. In: Data Analysis and Informatics, III, E. Diday, M. Jambu, L. Lébart, J. Pagés and R. Tomassone. Amsterdam, North Holland, 1984.

[23] OWSIŃSKI J., ZADROŻNY S. Structuring a regional problem: aggregation and clustering in orderings. *Appl. Stoch. Models and Data An.*, **2** (1986) 83–95.

[24] PARETO V. Manuel d'économie politique. Paris, Giard, 1927.

[25] REGNIER S. Sur quelques aspects mathématiques de problèmes de classification automatique. *I.C.C. Bull.*, **4**, (1965), 175–191.

[26] ROUBENS M. Mediane et méthodes multicritères ordinales. *Rev. Belge de Stat., Informatique et Rech. Op.*, **22** (1982), 2.

[27] RUBIN J. Optimal classification into groups: an approach for solving the taxonomy problems. *J. Theor. Biol.*, **15** (1967), 103–144.

* other than papers appearing in this volume.

[28] SPÄTH H. Cluster-Formation and — Analyse. Theorie, FORTRAN — Programme und Biespiele. Munich, Oldenbourg, 1983.
[29] TREMOLIÈRES R. The Percolation Method for an efficient grouping of data. *Pattern Recognition.* **11** (1979), 255–262.
[30] De la VEGA W. F. Techniques de classification automatique utilisant un indice de ressemblance. *Rev. Française de Sociologie.* (1967a).
[31] De la VEGA W. F. Quelques propriétés des hiérarchies de classifications. In: Archéologie et calculateurs, Colloque du CNRS, Marseille, (1976b).

Optymalizacja w analizie skupień: pewne podejście jako tło dla innych

Artykuł ten stanowi wprowadzenie do specjalnego numeru kwartalnika „Control and Cybernetics", poświęconego sformułowaniom zadań i podejściom do ich rozwiązywania odwołującym się do metod optymalizacji w zastosowaniu do analizy skupień. Artykuł przedstawia najpierw skrótowo pewne podejście oparte na optymalizacji dla funkcji celu implikującej rozwiązanie globalne w przestrzeni podziałów, a następnie pobieżnie wprowadza w tematykę tego numeru ukazując sformułowania i dziedziny zastosowań pojawiające się w poszczególnych prezentowanych tu pracach.

Оптимизация в кластерном анализе: некоторый подход как фон для других подходов

Статья является введением для особенного номера ежеквартального журнала Control and Cybernetics. Этот номер посвящен формулировке и методам решения задач, в которых употребляются методы оптимизации в приложении к кластерному анализу.

В статьи вкратце предсавлен некоторый подход использующий целевую функцию — критерий — имплицирующую глобально-оптимальное решение в пространстве разделений — совокупности кластеров, а затем обсуждены в общих чертах эти формулировки задач и применения, которые находятся в других статьях номера.