Control and Cybernetics

vol. **24** (1995) No. 4

Clustering - modelling, capacities, limits, applications

by

Jan W. Owsiński

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

This short introductory paper indicates the areas of concern in cluster analysis. First, it concentrates on the very definition of the domain of cluster analysis and on the adequacy of the methods developed therein. Then, it proposes a general approach to both modelling and resolution of the clustering problem, forming a framework into which several of the existing methods can be also accomodated. The thus outlined domain of cluster analysis is therafter characterized as to its limits, and some areas of applications are also indicated.

1. A foreword

Ten years ago – or by just a fraction less – a special issue of *Control and Cybernetics* was published, devoted to **Optimization Approaches in Cluster Analysis** (*Control...*, 1986). Publication of this special issue was equally motivated by the personal interest of the then – and now – guest editor and by a more general recognition of the importance of this field in data analysis. Since that time some fashions went away and some new came, but the essential problems remained unsolved. Thus, we still see a *proliferation of applications*, often done with ad hoc (new) methods or ad hoc modifications of the known ones. It seems, though, that no truly new methods have emerged over the last dozen or so years. Then, there is still lack of a more formal and constructive *junction with the methods of mathematical statistics* (e.g. inadequate effectiveness, but also serious formal shortcomings of these statistical methods which would at least verify the appropriateness of the number of clusters formed). Finally, largely in connection with the previous, cluster analysis is not being recognized as a *well-founded discipline* or – even less – a *theory* within data analysis or broadly conceived statistics. It is seen, instead, as a collection of approaches, methods and algorithms which try to do a similar, but by no means the same thing, and that in quite a variety of ways, most of which find a weak theoretical justification. Thus, cluster analysis is rather treated as a set of techniques or an

art to be mastered in order to deal with a certain poorly defined class of situations. This is largely due, in turn, to lack of the generally recognized – even just within the community of those dealing with cluster analysis – *formulation of the central problem of clustering*, the methods and potential solutions put aside. This is why the present author, and the guest editor, seizes this opportunity to return to the question of general formulation.

2. What is cluster analysis?

Cluster analysis is the discipline which deals with precise formulation, modelling, study and solution methods for the following problem, denoted further on as (1):

> having n objects indexed i, $i \in I = \{1, \ldots, n\}$, characterized by vectors $x_i = [x_{i1}, x_{i2}, \ldots x_{im}]$, $k \in K^x = \{1, \ldots, m\}$, and/or by distances d_{ij}^k, $i, j \in I$, $k \in K^d = \{m+1, \ldots, m^d\}$, and/or by proximities s_{ij}^k, $i, j \in I$, $k \in K^s = \{m^d + 1, \ldots, m^s\}$, to find such partition(s) P of I, $P = \{A_q\}_{q=1}^{p(P)}$, $A_q \subseteq I$, $\bigcup_{q=1}^{p(P)} = I$, that objects belonging to the same A_q's of this P be possibly close to each other, while those belonging to different A_q's be possibly distant. (1)

(We will refer further on to A_q resulting from a solution to (1) as to clusters.)

This formulation is sometimes – though this is by no means necessary – complemented with the condition on P that A_q's be mutually disjoint, $A_q \cap A_{q'} = \emptyset$, $q \neq q'$. Other details, like turning of (x_i, x_j) into distances and/or proximities, aggregation with/of d_{ij}^k and/or s_{ij}^k etc. will be omitted entirely in this short introduction as irrelevant for the main question.

The main question, namely, refers to this – crucial – fragment of (1) stipulating that "objects belonging to the same clusters be possibly close, while those belonging to different clusters – possibly distant".

Let us first note that this, quite obviously, implies existence of – or even is equivalent to – an objective function which is maximized over all feasible P and is composed of two elements:

$$Q_S^D(P) = Q^D(P) + Q_S(P) \tag{2}$$

of which $Q^D(P)$ represents the differentiation (distancing) of objects among various clusters, measured over the whole partition P, while $Q_S(P)$ represents the similarity of objects contained in the same clusters, as also measured over the whole partition P. The choice of sum, though arbitrary, does not seem to limit the generality of (2) nor its adequacy to (1).

Let us also note that we may quite as well deal with a "dual" to (2), namely

$$Q_D^S(P) = Q_D(P) + Q^S(P) \tag{3}$$

with the principles of interpretation preserved from (2). This "dual" is, of course, minimized, and it is just as adequate representation of (1) as (2) is.

Thus, we have a good expression for (1), but what next? What comes next is, first of all, the statement that if (2) is correct, then a vast majority of existing clustering schemes (agglomerative, divisive, K-means, ...) cannot be considered as representing appropriately (1) and, therefore, as leading to its solution. Then, we can postulate a number of obvious properties of $Q^D(P)$ and $Q_S(P)$ and try to complement at least these of the existing schemes which are based upon just one of the two components with the other one, while attempting to preserve (the good) computational properties of these schemes.

On the other hand, the previous papers by the present author (Owsiński, 1980, 1984, 1990) indicate the possibility of formulating a very general suboptimization procedure, whose main assumption is that $Q^D(P)$ and $Q_S(P)$ have opposite monotonicity along the mergers/splits of $A_q \in P$. This property is satisfied by numerous functions which can be used as adequate representations of (1). Thereby, not only a family of easily suboptimizable $Q^D_S(P)$ and/or $Q^S_D(P)$ can be obtained – see Owsiński (1990 and 1992, with some 30 examples of such functions in the latter), but also, as suggested before, the already existing schemes can in many cases be adapted so as to attain conformity with (2) or (3) and preserve appropriate computational facility.

3. The procedure

The suboptimization procedure for (2), based upon the opposite monotonicity assumption mentioned, takes, for instance, the following form for $Q^D_S(P)$:

Assume that $Q^D(P)$ increases with aggregation of A_q's, while $Q_S(P)$ decreases. Thus, if instead of (2) we take

$$Q^D_S(P,r) = (1-r)Q^D(P) + rQ_S(P) \rightarrow \max,\ r \in [0,1], \tag{4}$$

and start the procedure, with the step number $t = 0$, $r^0 = 1$, we obtain $P^0 = \text{argmax}Q^D_S(P,r^0) = \text{argmax}Q_S(P) = I$. This is the initial point of the procedure. Then, for given P^t, r^t, consider the partitions $P^t_H(q,q')$ which differ from P^t by aggregation of A_q and $A_{q'}$. Find, for each pair $(q,q') \in \sqcap(P^t) \times \sqcap(P^t)$, where $\sqcap(P)$ are the sets of indices q for definite partitions P, the value of r satisfying

$$(1-r)Q^D(P^t) + rQ_S(P^t) = (1-r)Q_D(P^t_H(q,q')) + rQ_S(P^t_H(q,q')) \tag{5}$$

i.e. the condition of aggregation for the particular pair of clusters $\in P^t$. This value of r is equal

$$r^t(q,q') = \frac{Q_D(P^t_H(q,q')) - Q_D(P^t)}{Q_D(P^t_H(q,q')) - Q_D(P^t) + Q_S(P^t) - Q_S(P^t_H(q,q'))} \tag{6}$$

which means that $r^t(q, q') \in [0, 1]$ indeed. Since we are interested in the (locally, of course, from the point of view of the whole procedure) best aggregation, we set

$$r^{t+1} = \max_{q,q') \in \sqcap(P^t) \times \sqcap(P^t)} = r^t(q^*, q^{**}) \tag{7}$$

and accept the respective $P_H^t(q^*, q^{**})$ as P^{t+1}.

This procedure retains as suboptimal the partitions P^t for which either $0.5 \in [r^t, r^{t+1}]$ or $0.5 \in [r^{t-1}, r^t]$, in view of the form of (2).

4. The limits of clustering

On the basis of the approach outlined, composed of the objective function (2) or (3) meant to model (1), and the simple hierarchical merger procedure (4) through (7), one can not only obtain the suboptimal solutions to the clustering problem (1), but also several numerical characterizations of this solution (values of $Q^D(P^t)$, $Q_S(P^t)$ and r^t). Even, though, with such a procedure, which can additionally be complemented with certain improvement schemes, there are some essential limitations to the clustering approach as represented by (1). These limitations are primarily related to the interpretation of the partitions obtained. If we know *a priori* sufficiently well what is the meaning of our $Q^D(P)$ and $Q_S(P)$ then we are certain as to the "appropriateness" of the solutions obtained. This, however, is very rarely the case. And although the opposite is also true (i.e. any criticism as to the results of clustering has to explicitly refer to $Q^D(P)$ and/or $Q_S(P)$), we are often confronted with the problem of analysis of "meaning" of the partitions resulting from the procedure(s). This analysis goes in two main directions:

(a) the testing of statistical "objectivity", and
(b) the verification of the adequacy for a given practical application.

Both these kinds of issues are linked with the degree of ability of turning the objective functions (2) or (3) into the respective counterparts: appropriate statistics or objective functions of a given applied problem. This ability, as of now, is very limited, indeed, if at all any (Gordon, 1995). It is true that with respect to the juncture with statistics the paper by Marcotorchino (1986) shows how one of possible embodiments of (2) can be parallelled with a number of classical statistics, but this is only for just one embodiment and, moreover, the constructive nature of the parallels is very limited.

Indeed, the (only) constructive approaches to clustering, which come from the statistical domain, referring to the mixture model or to the Bayesian estimators, often end up – if formulated in practicable terms – with quite cumbersome optimization problems in which, though, only one side of (2) is accounted for (the "loss function", for instance). Introduction of a formally statistically based complement to this kind of function, although intuitively relatively obvious,

would require a lot of theoretical effort and the concrete forms are as yet not in sight.

5. Applications

The latter difficulty can also well be illustrated by the cases of applications, especially, but not exclusively, the ones which appear recently. We will concentrate here on just one area of applications, namely those related to "flexible manufacturing" or "group technology". The respective references may be provided by, for instance, Kusiak (1992), Chen (1993), Srinivasan (1994) or Ben-Arieh and Chang (1994).

The original problem which is being solved in the publications referred to is a concrete production planning or scheduling problem, for which definite mathematical models can be built, usually taking the form of the optimization tasks. These optimization tasks, though, both in view of their inherent complexity (integer programming, nonlinearity, multiplicity of various constraints, dynamic nature) and of the dimensions (e.g. numbers of products, time instants, machines, ...), turn out to be intractable in practice. Thus, heuristic approaches are devised that will provide approximate solutions to such problems. In many cases these heuristic approaches take the form of the clustering problem, generally equivalent to (1). This is related to the fact that the notions of distance and/or proximity (likeness) as well as global quality of partition are somehow adequate to the original problem formulation (e.g. similar characteristics of products from the point of view of their production process, coupled with the necessity of assigning them to a number of distinct production lines/centres).

It must be emphasized, though, that there is, as of now, virtually no direct formal link, in terms of a transformation or at least a well-founded bound, between the original formulation of the optimization problem and the one of clustering. In fact, in numerous cases the clustering heuristics applied do not even refer to explicit formulations of a model of (1), be it in the form of (2, 3) or any other appropriate one, but simply use the existing known techniques, of which we have already said that they do even not address adequately the problem (1).

Thus, in the domain of applications the need exists of finding a way to model some important classes of practical problems through clustering formulations. It seems that (2, 3) provides a framework for such a way. Likewise, it may be hoped that this will turn out also to be the proper way for making the connection with the statistical domain, as indicated already, though quite narrowly, by Marcotorchino (1986).

6. On this volume

The present issue of *Control and Cybernetics* shows quite a range of work which is being currently done in the clustering domain: from statistical models of rel-

atively complex nature, through new developments in classical clustering techniques (single link) and in fuzzy-set based methods, down to very practical applications of known algorithms implemented in widely disseminated packages.

Thereby, a proper illustration for the image of the domain is provided, together with the technical and more in-depth questions which are being solved.

We start with a very interesting paper by Yu. Kharin and E. Zhuk, which considers the Bayesian classification model with Markov-type dependence between the class number assignments. This paper is then followed by two papers (by K. Jajuga and by S. Miyamoto and Y. Agusta) dealing with fuzzy clustering methods. The sequence chosen is primarily related to the perhaps subtle, but still very important junction between the statistical models and the fuzzy set K-means ones, which is provided by the notion of loss function, representing indeed one side of (1). Attention should be paid to the new result shown in the second of the two papers mentioned.

The subsequent group of papers deals with the efforts aiming at finding of effective and efficient methods for optimization in case of clustering problems, though not always of the very general nature of (1). Thus, G. Govaert considers simultaneous clustering of rows and columns, in fact one of the oldest specific problems in cluster analysis, and proposes the method(s) which result from quite a broad overview of the domain. Then, P. Kadłuczka and K. Wala on the one hand, and J.S. Chipman and P. Winker on the other propose heuristic methods largely referring to the so called "stochastic" vein in optimization to solve definite clustering problems. It is interesting to note, especially with respect to the paper by J.S. Chipman and P. Winker, that quite practical problem is being solved with a far reaching degree of detail. Finally, Th. Gafner returns to the old question of applicability of dynamic programming to clustering and shows the ways in which the approach left already some time ago as apparently inefficient can be improved and therefore perhaps used in some cases when exact solutions are needed.

To somehow close the landscape of cluster analysis we have, at the end, first the paper by Ph. Lehert and Ch. Dumortier which presents a new technique for the single link method which, given certain assumptions, can go down with its computational complexity to $O(n)$, though at the expense of rapid increase of computational effort with m, the number of variables describing x_i. Still, the result shown is of particular significance in view of the often encountered difficulty in dealing with very large sets of data (e.g. when $n = 10^5$ or even 10^6). The last paper in the volume, by E. Kovacs amd A. Sugar, shows a practical application of a straightforward clustering technique "taken from the shelf". This is this end of the domain which we all ought to have well in mind when devising new approaches, methods and techniques.

References

BEN-ARIEH D. AND CHANG P.T. (1994) An extension of the p-median group

technology algorithm. *Computers & Ops.Res.*, **21**, 2, 119-125.

CHEN C.Y. (1993) Cluster first – sequence last heuristics for generating block diagonal forms for a machine-part matrix. *IJPR*, **31**, 11, 2623-2647.

Control and Cybernetics (1986): *Optimisation Approaches in Clustering*, guest editor: Jan W. Owsiński, **15**, 2.

GORDON A. D. (1995) Determining the number of clusters. Paper presented at the Third Meeting of the French-Speaking Classification Society. Namur, Belgium, September 28-29, 1995.

KUSIAK A., ED. (1992) *Intelligent Design and Manufacturing.* John Wiley, New York.

MARCOTORCHINO F. (1986) Cross association measures and optimal clustering. *Compstat '86*, F. de Antoni, N.Lauro and A.Rizzi, eds. Physica-Springer, Heidelberg.

OWSIŃSKI J.W. (1980) Regionalization revisited: an explicit optimization approach. CP-80-26, IIASA, Laxenburg (Austria).

OWSIŃSKI J.W. (1984) On a quasi-objective global clustering method. In: *Data Analysis and Informatics*, E.Diday et al., eds. North Holland, Amsterdam.

OWSIŃSKI J.W. (1990) On a new naturally indexed quick clustering method with a global objective function. *Applied Stochastic Modelling and Data Analysis*, **6**, 157-171.

OWSIŃSKI J.W. (1991) Nowa metoda analizy skupień z globalną funkcją celu (New method of cluster analysis with a global objective function; in Polish). Ph.D. dissertation. Systems Research Institute, Polish Academy of Sciences.

SRINIVASAN G. (1994) A clustering algorithm for machine cell formation in group technology using minimum spanning tree. *IJPR*, **32**, 9, 2149-2158.