ON A TWO-STAGE CLUSTERING PROCEDURE AND THE CHOICE OF OBJECTIVE FUNCTION

J.W. Owsiński
Systems Research Institute, Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, POLAND
e-mail: owsinski@ibspan.waw.pl

Abstract

The paper introduces a class of simple hybrid clustering algorithms, based on the idea of obtaining, first, a high number of "subassemblies" or "cells", through application of a k-means-type procedure, and, second, aggregating these "subassemblies" into shapes with a hierarchical merger procedure. For n objects, characterized by vectors of values x_i, $i = 1, , n$, we look in the first stage for p_1 cluster "cells", with $n >> p_1 >> 1$, possibly $p_1 \sim O(n^{1/2})$. Then, distances between the thus formed clusters, $A^*_{q_1}, q_1 = 1, , p_1$, are calculated according to one of the predefined formulas. On the basis of these distances the second stage algorithm is executed of the hierarchical merger kind. The final number of output clusters A_{q_2}, i.e. $p_2, p_1 >> p_2 \geq 1$, is determined with the use of the global objective function for the clustering, developed by the author. The generic method is primarily meant to recover the clusters of cumbersome, curvilinear, shapes, which it does effectively and efficiently. The class of algorithms is generated by (a) choice of a particular k-means-type algorithm for the first stage; (b) choice of a particular distance measure $d(A^*_{q_1}, A^*_{q_1})$; (c) choice of the hierarchical merger algorithm, coupled with the concrete form of the objective function. The overall algorithms thus obtained differ significantly by their efficiency, numerical complexity and effectiveness, as well as the nature of output clusters.

Key words: clustering, k-means, hierarchical merger, hybrid algorithms, curvilinear clusters, objective functions

1 The motivation and the general outline

In the efforts to enhance the effectiveness of existing clustering techniques, a two-stage strategy, associating the advantages of two different kinds of algorithms, better fit for the initial and the final stages of clustering, appears as quite appealing. The best known example of such an approach seems to be "The SPSS TwoStep Cluster Component" [5]. Here, another two-stage approach is presented, stemming from different premises and purposes. Thus, while effectiveness in treating (relatively) large sets of data was definitely one of the aims, it was deemed far more important to (1) be able to possibly easily recover clusters of awkward, curvilinear shapes and (highly) variable magnitude characteristics, and (2) ensure flexibility with respect to the assumed "parameters", guiding the solution finding process. With this respect the choice was relatively simple: first a k-means type algorithm, over which a user can have quite "tight" control in many respects, and then a selection of the agglomerative clustering algorithms. Flexibility is

not only assured by the possibility of choosing among the available techniques at both stages, but also the options both at the beginning of the procedure (starting point) and the junction of the two stages.

A separate question is, as usual, constituted by finding of the "right" number of clusters. For this purpose the general global objective function, developed by the author, was used. This choice results not only from the conviction of the appropriateness of this approach, but, primarily, from the fact important in view of the motivations mentioned that this function lends itself to various formulations that may be fitted to the selected agglomerative clustering algorithm.

2 The problem and the procedure

Given n objects indexed i, $i \in 1, ..., n = I$, each described by a vector x_i of values x_{ik}, $k = 1, ..., m$, we look for the best clustering of I into clusters $A_q, q = 1, ..., p$, where p (or $p(P)$, where P is a partition) is the a priori not defined number of clusters. We can postulate that variables indexed k form the space of all potential objects, E_X, so that if all x_i form the set X_I, then $X_I \subseteq E_X$. We refer to the essential formulation of the clustering problem, i.e. "find division of I into clusters such that objects in the same cluster are possibly close, while objects in different clusters possibly distant". For this, we only assume that both distances $d(.,.)$ and similarities $s(.,.)$ can be appropriately calculated (denoted d_{ij} or s_{ij} for the pairs of objects in I). Of A_q we assume only that they sum to I, although respective algorithms may produce non-overlapping clusters, i.e. $A_q \cap A_q = \oslash, q \neq q$. Whenever applicable, the representative object of a cluster, whether belonging to X_I, or to $E_X - X_I$, will be denoted x^q (it is assumed that there is only one such object per cluster).

The procedure is divided into two stages: In the first stage an algorithm of the k-means family is performed with a predefined number of clusters, p_1 (users choice), selected at a relatively "high" level much higher than the expected "ultimate" ("objective"?) number of clusters. Just as a hint, for a wide range of values of n one can use $p_1 = n^{1/2}$. So, clusters A_q^1 are obtained, $q = 1, ..., p_1$.

Once the first stage terminated, the matrix of distances between clusters A_q^1 is calculated, D^1. On the basis of this matrix, one of the classical progressive merger algorithms is performed, the default choice being the single link (nearest neighbour) procedure.

The working of the agglomerative clustering scheme is accompanied by calculation of values of the global objective function of the author, [2, 3], leading to determination of the sub-optimal number of clusters (see Section 4), so that, ultimately, the aggregation process can stop before reaching $p = 1$.

3 The choices offered

Since the first application was ready in early 2007, see [4], the technique is still in the experimental phase, primarily from the point of view of selection of the "best-fitted"

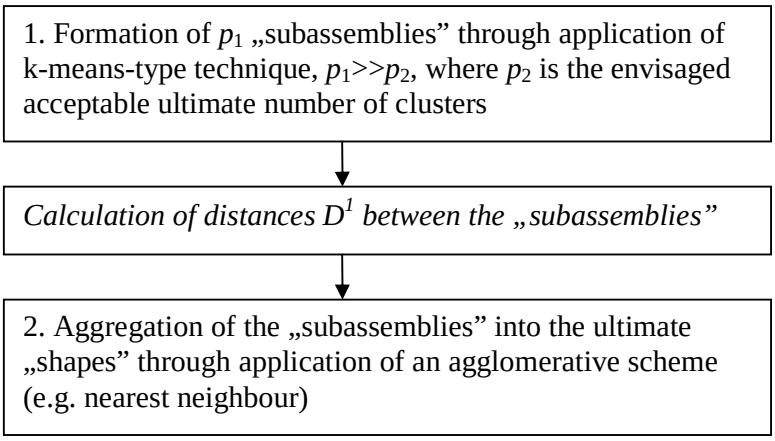

Figure 1: The scheme of the two-stage algorithm

option paths along the procedure. The options offered, conform to the logic of the hybrid procedure, are:

1. generation of the starting point for the k-means-type algorithm: notwithstanding the obvious differences, resulting from the selection of k-means or k-medoids, it is possible to start with a grid, spanning the E_X space, or the X_I set, this choice being quite reasonable, provided the specified p_1 is high enough;

2. the first stage algorithm: now the choice is between standard k-means and k-medoids, but it is envisaged that FCM-type algorithm will be included, as well as specialized techniques for dealing with categorical variables (like k-histograms); use of other techniques, related to the E-M approach, is not excluded, either; there is also a possibility of pre-specifying the number of iterations of the procedure, given that we are interested only in the "subassemblies", which can be of quite rough character;

3. the distance definition: as of now the basic choices are available, referring to Minkowski distances and the standard ones for categorical data;

4. the number of clusters obtained from the first stage, p_1: the default is $n^{1/2}$, but a user can specify virtually any number, the only check is on the maximum not exceeding n; so, the entire procedure might boil down to k-means-like algorithm, if the specified number is appropriately small (see next section for the use of the objective function);

5. the way, in which distances D^1 are calculated between the clusters, obtained from the first stage; the basic choices with this respect include: (5.1) the distances between the centroids, or medoids; (5.2) the minimum distance between any objects from two clusters considered (a simile of the single link); (5.3) an estimate of the minimum distance, e.g. minimum distance between the 10% of objects farthest away in each cluster from the respective centroids or medoids; (5.4) the

maximum distance between any objects from two clusters (simile of complete link); let us note that the choice, made at this point, is very important for the numerical properties of the overall procedure, and also from the point of view of algorithmic consistency (i.e. whether D^1 matches the subsequent agglomerative algorithm);

6. the second stage algorithm: as of now choice can be made only among three basic schemes (single link, complete link and average link); there are no plans to essentially broaden this choice, perhaps only with one or two schemes (Ward included);

7. determination of the ultimate number of clusters: here the choice is now among three pre-specified forms of the global objective function, which is used to determine the cluster number (see next section).

Although some comparisons were performed with the SPSS TwoStep technique, their results are not presented here insofar as the actual comparability could not be established in view of the wide difference in options subject to selection on both sides. Generally, of course, time-wise the SPSS code performed much better, although in some cases had more problems with establishing the "right" number of clusters. In terms of time and memory requirements, the hybrid procedure depends largely upon the choices under 4 and 5 above. The code was not optimized, since the primary goal was associated with shape recovery, but it must be indicated that given the possibility of analyzing in the first stage a high number of small clusters, the number of k-means iterations can be brought to a minimum, approximating quite effectively the single-pass algorithms.

4 Determination of the cluster number with the global objective function

The global objective function was introduced in the early 1980s by the author, [2, 3], with the aim of determining the cluster content along with cluster number within one integral procedure rather than referring to an "external" criterion, not associated with the clustering procedure, to check whether the clustering(s) obtained (or which of them) satisfy it.

Strictly conform to the verbal formulation of the clustering problem, quoted before the method refers to an explicit objective function, being the function of quality of the partitions, denoted $Q_S^D(P)$, which is composed of two parts:

$$Q_S^D(P) = Q^D(P) + Q_S(P) \tag{1}$$

where $Q^D(P)$ denotes the function of distance between objects assigned to various clusters, defined over the whole space of partitions $E_P, P \in E_P$, while $Q_S(P)$ is the analogously defined (in terms of interpretation, and not necessarily the very form) function of similarities among all the objects assigned to the same clusters. The function

$Q_S^D(P)$ is maximized. Its maximization allows - at least in principle - to determine both the composition of the optimal A_q and the optimum p, number of clusters. This unique feature is the consequence of the *global* nature of the general objective function (1).

The *generality* of (1) is reflected through the possibility of accommodating a broad variety of concrete formulations within the framework proposed. First, instead of maximizing (1) we can minimize its "dual", in which $Q_D(P)$ is the function of distances between the objects within the same clusters (like, for instance, in the k-means-like techniques), while $Q^S(P)$ is the function of similarities between objects in different clusters. Then, the way in which the functions are formulated is subject to choice, driven by the nature of the problem at hand, and the numerical facility of resulting computations.

The approach assumes simultaneous use of distances $d(.,.)$ and similarities $s(.,.)$ between objects, d and s being linked by some simple and obvious functional dependence, $s(d)$ and $d(s)$, and so, ultimately, one might deal away with the explicit use of both notions. Still, for the sake of clarity of interpretation, we stick to the explicit use of d and s.

Yet, according to the method associated with the function, which includes also an algorithm of the search for the solution, the concrete formulations of the function (1), or its "dual", are subject to a condition of algorithmic nature. In the framework of the approach, namely, for algorithmic purposes, (1) is transformed into a parametric form

$$Q_S^D(P,r) = rQ^D(P) + (1-r)Q_S(P) \tag{2}$$

$r \in [0,1]$. With this general parametric form, the function is suboptimized through a simple, classical progressive merger procedure. The procedure starts with $r = 1$, when the parametric function is equivalent to $Q^D(P)$, whose maximization yields the optimum partition $P^*(1) = I$, meaning that each object forms a separate cluster. As the value of r decreases, consecutive partitions $P^*(r)$ are formed through mergers of clusters created at earlier stages.

In order for this procedure to lead to (sub)optimum solution (for $r = 0.5$), the condition of *opposite monotonicity with respect to aggregation/disaggregation of clusters* is applied to the components of $Q_S^D(P)$. It means that when we obtain from a given partition P another partition through a merger of arbitrary clusters into one, then the resulting changes in the values of $Q^D(P)$ and $Q_S(P)$ should go in the opposite directions, and with every possible aggregation these directions will for a given component always be the same. An analogous principle is valid for the disaggregation (dissection) of clusters. It turns out that there exists a broad class of concrete forms of $Q^D(P)$ and $Q_S(P)$, which fulfill this condition, see [3].

By application of functions satisfying "opposite monotonicity", the obtained progressive merger scheme, similar to those of the Lance-Williams-Jambu (L-W-J) formula, yields sub-optimum solutions. The scheme is generally as effective as the L-W-J ones and suboptimizes both the *contents of clusters and their number*. The actual effectiveness of algorithms will depend, of course, on the form of the concrete functions $Q^D(P)$ and $Q_S(P)$. Additionally, the algorithm is equipped with a *natural index* r of

hierarchy, whose values, obtained during the functioning of the algorithm, may serve to assess the robustness of the individual partitions forming hierarchical structure.

The examples of concrete forms of (1), which are, anyway, used in the present version of the hybrid algorithm, are:

$$Q_S^D = Q_S + Q^D = \Sigma_{q\in C(P)}\Sigma_{i,j\in Aq} s_{ij} + \Sigma_{q,q'\in C(P)}\Sigma_{i\in Aq,j\in Aq'} d_{ij} \tag{3}$$

$$Q_S = \Sigma_q \Sigma_{i\in Aq} \max_{j\in Aq, j\neq i} s_{ij} \text{ and } \mathrm{Q^D} = \Sigma_{\mathrm{q}}\Sigma_{\mathrm{q'>q}} \min_{\mathrm{i\in Aq, j\in Aq'}} \mathrm{d_{ij}}, \tag{4}$$

$$Q_S = \Sigma_q \Sigma_{i\in Aq} \min_{j\in Aq, j\neq i} s_{ij} \text{ and } \mathrm{Q^D} = \Sigma_{\mathrm{q}}\Sigma_{\mathrm{q'>q}} \max_{\mathrm{i\in Aq, j\in Aq'}} \mathrm{d_{ij}}, \tag{5}$$

where $C(P) = \{1, ..., p(P)\}$, these three formulations corresponding in a certain manner to three basic agglomerative schemes: average, single and complete link, respectively. While, however, (3) provides for a strict analogue of the average link (see de Falguerolles, 1977), (4) and (5) are just similes of the, respectively, single and complete link. For this reason, in particular, in the current implementation of the hybrid algorithm not the suboptimisation procedure, outlined before, based on (3-5), is used, but the corresponding classical procedures, whose output is only evaluated with these objective functions.

5 Some results

The algorithm fared very well with a number of exemplary instances, designed to test the capacities, for which it was designed, i.e., first of all recovery of cumbersome shapes. Fig. 2 shows the two-dimensional case, which is representative for this kind of problems treated. Then, Fig. 3 shows a simple case, illustrating the examples meant primarily to test the properties of the objective functions here quoted (most of these examples were much bigger in terms of n and usually involved a three-level organization of objects).

The case of Fig. 2, when treated with the algorithm at $p^1 = 50 - 60$, yielded correct results for pre-defined $p = 6$ (while, of course, k-means nor k-medoids could do this, chopping the "objective" clusters into pieces). Application of the objective function (4) implied the choice of $p = 5$ (maximum value), obviously suggesting that the ring and its centre should constitute one cluster. It must be admitted, though, that the difference in the value of the objective function between $p = 5$ and $p = 6$ is virtually negligible (at the order of 0.1% of the entire range of values of (4) for this example), while differences with partitions obtained for other values of p (preceding and subsequent steps of the procedure) are much bigger.

The cases like that of Fig. 3 yielded local maxima or plateaus of the objective function for cluster numbers, corresponding to respective "levels of organization" of the objects (these cases, though, were treated with complete linkage and the function (5)).

In Fig. 4 an exemplary course of the objective function (3) is shown for quite a complex set of test data ($n > 1000$, $m = 6$), derived from a mixture of transformed normal populations, presented by G. Ritter [6]. In an obvious manner, the objective

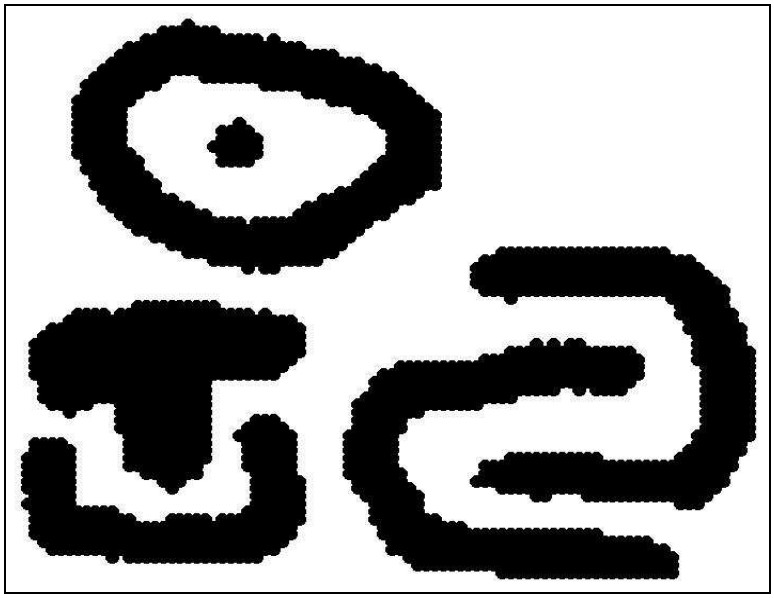

Figure 2: A "clinical" two-dimensional example treated, with $n = 2277$.

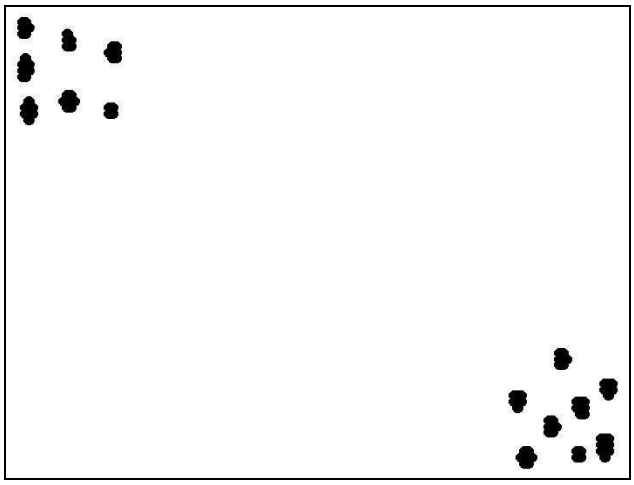

Figure 3: A simple "two-level" example treated, with $n = 106$.

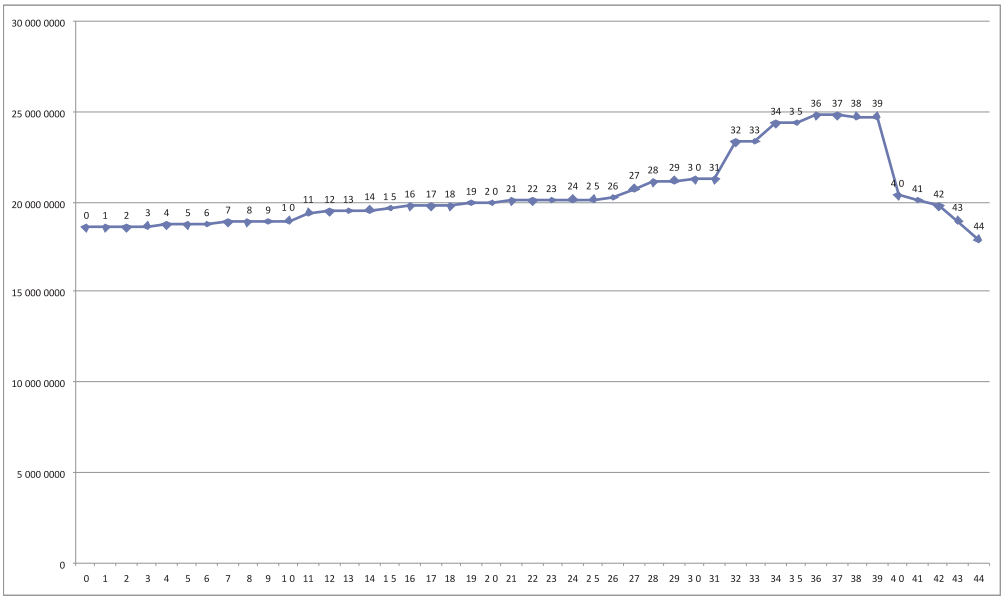

Figure 4: An example of the course of the objective function (3) for the test data of G. Ritter [6]

function indicates the "optimum" number of clusters, or a sequence of "near-optimum" cluster numbers, as well as those much farther away from the optimum.

6 Conclusions and future work

The hybrid scheme proved to be both effective and flexible. Flexibility is assured by the implemented set of options, concerning both the two stages of the procedure, and, especially, the linking element of the distance calculation. Although no match for the SPSS TwoStep Cluster Component in terms of time, it performs relatively well with respect to both time and memory requirements under a simplified set of options. What is even more important from the point of view of the initial motivation is that it identifies well shapes, both composed of "thin lines" and of "thick lines". The fact of using the global objective function for evaluation of partitions obtained from the second stage, which can be tuned to a particular agglomerative algorithm, definitely adds to the quality of results (few partitions can be picked, characterized by the extreme values of the objective function, and compared with respect to other features, "external" to the procedure). An additional capacity is provided by the possibility of "re-aligning" with the objective function, proper for the first stage.

The entire procedure shall be further developed in two directions: (1) more flexibility in terms of choices, mentioned in Section 3; (2) choice of best paths, formed by these choices (and respective hints to the user). This second direction of work shall be closely associated with thorough testing and comparison, both with the directly

"competitive" SPSS technique, the algorithms used in their classical versions, and, of course, among various option paths of the hybrid approach.

References

[1] H. Daschiel, M.P. Datcu. Cluster structure evaluation of dyadic k-means for mining large image archives. In: B. Serpico, ed., *Image and Signal Processing for Remote Sensing VIII*. Proceedings of the SPIE, **4885**, 120-130, 2003.

[2] A. de Falguerolles. Classification automatique: un critere et des algorithmes d'change. In : E. Diday and Y. Lechevallier, eds., *Classification automatique et perception par ordinateur*. IRIA, Le Chesnay, 1977.

[3] J.W. Owsiński. On a quasi-objective global clustering method. In: Diday, E., Jambu, M., Lbart, L., Pags, J., Tomassone, R., eds., *Data Analysis and Informatics III*. North Holland, Amsterdam, 293-306, 1984.

[4] J.W. Owsiński. On a new naturally indexed quick clustering method with a global objective function. *Applied Stochastic Models and Data Analysis*, 6, 1990.

[5] J.W. Owsiński, M.T. Mejza. On a New Hybrid Clustering Method for General Purpose Use and Pattern Recognition. In: *Proceedings of the International Multiconference on Computer Science and Information Technology*, **2**, ISSN 1896-7094, http://www.papers2007.imcsit.org/ 121-126, 2007.

[6] G. Ritter. the test data sets presented at the 1^{st} German-Polish Workshop on Data Analysis, Aachen, October 2009.

[7] SPSS: The SPSS TwoStep Cluster Component; retrieved from www.spss.com as late as March 9^{th}, 2009.

[8] T. Zhang, R. Ramakrishnon and M. Livny. BIRCH: an efficient data clustering method for very large databases. *Proc. of the ACM SIGMOD Conference of Management of Data*, 103-114, 1996.