



A Hybrid Clustering Method for General Purpose and Pattern Recognition

Jan W. Owsinski¹, and Mariusz Tomasz Mejza²

¹ Systems Research Institute, Polish Academy of Sciences, Newelska 6,
01 447 Warsaw, Poland
owsinski@ibspan.waw.pl

² Warsaw School of Information Technology, Newelska 6,
01 447 Warsaw, Poland
mejza@wsisiz.edu.pl

Abstract. The present short paper outlines a simple hybrid technique of clustering, based on the joint use of k-means and nearest neighbour algorithms. The technique starts with the k-means algorithm, performed as the first stage for an adequately high number of centroids and continues with the nearest neighbour algorithm, executed for the clusters obtained in the first stage, as the set of initial objects to be merged. It is shown on examples how the thus defined technique performs in terms of identification of relatively complex shapes.

1 Introduction

The short note is devoted to presentation of a clustering technique, which is meant to overcome some of the shortcomings of the techniques existing as of now, while retaining their strong points, known from literature. The Introduction section will highlight the premises behind and the purposes of development of the new technique.

1.1 The Search for a More Flexible Technique

Clustering, i.e. *placing the similar together and the dissimilar apart*, is not only a model of the basic intellectual activity, and not only a fundamental problem in multivariate analysis, but, first and foremost, a tool used in multiplicity of domains. Although the general clustering problem can be considered to have found an ultimate solution – if at all – through the formulations like those of [1, 3] and [4], the search for more powerful (in terms of finding “true solutions”), more adapted (to numerous specific situations) and more efficient (in terms of computational effort) techniques is still on. Due to this, and due to the varied properties of the techniques proposed, the domain of cluster analysis is still developing, also in its theoretical aspect, as witnessed, e.g., by books such as [2].

The dozens of existing approaches have each some merits in one or more of the fields mentioned, but also display, inevitably, poor performance with respect to some other ones. This, again, propels the development, especially of the narrowly designed techniques, like those specializing in definite tasks of pattern recognition, document retrieval etc. Indeed, the technique, which we propose in the present paper, goes

along these lines, while at the same time preserving some decent level of generality. The ultimate goal, of course, is a flexible technique that could be adapted to various application cases.

1.2 The Reasons for the New Development

The reasons for the development of the new technique can be summarized as follows: (1) to develop a method of clustering that providing effective means for visualization; then (2) to thereby develop a method that can be effectively used in pattern recognition; and (3) to improve on the existing techniques in more general terms. These reasons acted in the sequence as here provided: from a more modest goal to a much more ambitious one. It was, namely, hoped that effective visualization would constitute a good starting point to the other two goals, but even if it were to stop there, the exercise would be worth the effort.

It must be added that while aiming at the technique aiding in visualisation of the data sets we planned to have, at this stage of work, a simple instrument that would be tested on several cases through human verification. Actually, this is also what visualisation is about.

2 The Technique

We shall now give the complete description of the approach, with just a short consideration of the technical details, which are left to ampler future publications on the subject, linked with further developments of the method.

2.1 Notation

Assume we deal with n objects (observations, items), indexed i , $i \in I = \{1, \dots, n\}$. Each object is described with m variables (attributes, features), of any character, and such description is denoted x_i , $x_i \in X_i$. We can postulate that these variables form the space of all potential objects, denoted E_X , $X_i \subseteq E_X$. Assume, further, that we can define a distance in E_X , with distance between objects considered denoted $d(x_i, x_j) = d_{ij}$. Distances d_{ij} form a symmetric matrix $D = \{d_{ij}\}_{ij}$.

The set I is divided (partitioned) in the clustering problem into subsets (clusters) denoted A_q , $q = 1, \dots, p$, where p is the number of clusters, forming a partition P .

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).

2.2 General Scheme of the Algorithm

The algorithm is divided into two stages:

In the first stage the k-means algorithm is performed with predefined number of clusters, p_1 (user's choice) at a relatively "high" level, anyway – 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 $p_1 = n^{1/2}$. In this manner, clusters A^1_q are obtained, $q=1, \dots, p_1$.

Once the first stage terminated, the matrix of distances between clusters A^1_q is calculated, D^1 . On the basis of this matrix the classical progressive merger procedure of single link (nearest neighbour) is performed.

Just like p_1 , the number of clusters ultimately determined might be an explicit choice of the user, p_2 , or the merger procedure can be carried out to the very end, with p_2 determined on the basis of additional information.

2.3 The Rationale

The use of the two known algorithms in the way here proposed is justified by the following reasoning: It is well known that k-means form spherical or ellipsoidal clusters and converge very quickly. If the clusters formed are “small”, and “dense” in the set of objects considered, the convergence is even quicker, and there is little hazard of finding a local minimum, so that either only few repetitions are needed, or they can be given up at all. By applying k-means in this way we obtain an effective breakdown of the data set into small, compact subsets, even though these subsets may have very little to do with the actual “shape” of the proper clusters sought. In the second stage single link is used, which agglomerates the subsets according to minimum distances, so that if these subsets form any linear and complex shape, it should get uncovered, with the single link algorithm not so much penalised by the necessity of maintaining and re-calculating the distance matrix, owing to the shrinking of the dimension of the problem in the first stage.

2.4 Some Technical Remarks

The primary issues of technical nature, which arise in the implementation and running of the algorithm are quite obvious:

- i. determination of p_1 : besides the hint provided above, caution must be made of the maintenance of the reasonable (?) proportions between n , p_1 and the envisaged p_2 ; one might also use a constant divisor, bringing n down to p_2 ; this issue is, of course, closely associated with the fact that neither k-means nor single link by themselves provide a way to determine the “proper” p_2 ;
- ii. generation of the initial centroid candidates for the k-means stage: given that we start with a much bigger number of centroids than the sought number of final clusters, the initial centroid candidates can be determined by a method different from random choice in E_X (or X_i);
- iii. calculation of the distance matrix D^1 ; this is the key issue in the computational efficiency of the algorithm; in the application developed to implement the method, a user is offered three options at this point: **(1)** complete enumeration (i.e. $d(A^1_q, A^1_{q'}) = \min \{d_{ij} : x_i \in A^1_q, x_j \in A^1_{q'}\}$) is obtained on the basis of all pairs i, j such that $x_i \in A^1_q, x_j \in A^1_{q'}$; **(2)** the value of $d(A^1_q, A^1_{q'})$ is calculated as the d_{ij} between $x_i \in A^1_q$ that is the closest to $x^{q'}$ and $x_j \in A^1_{q'}$ that is the closest to x^q ; **(3)** a predefined proportion (user’s

choice) of objects in both clusters is compared conform to the scheme (2) above.

The fact that these choices are offered in determination of D^1 comes from the contribution of this phase of functioning of the algorithm to the overall computational burden, both in terms of time and memory requirements.

3 Examples

This section presents two characteristic examples of functioning of the algorithm, selected so as to show the merits of the technique and to compare it with the classical k-means.

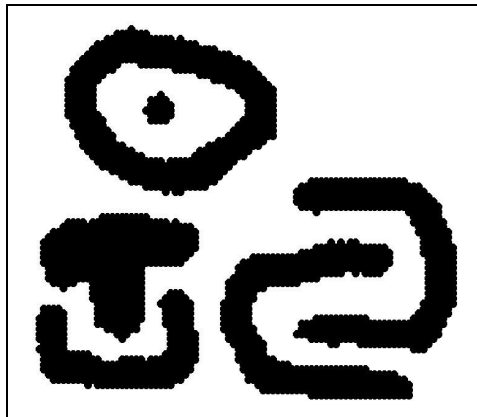


Fig. 1. A „clinical” case treated by the algorithm proposed, with $n = 2277$, $m = 2$

For the example shown schematically in Fig. 1 above, with relatively complex shapes of the clusters “to be identified”, the new algorithm, with $p_1 = 227$ and distances in D^1 calculated according to option (2), allowed to identify precisely the visually obvious clusters on the basis of the aggregation distance diagram. The classical k-means, for which $p^2 = 6$ was set, was unable to produce these clusters within a reasonable number of repetitions.

The other case treated, which is shown here, had a different objective. As shown in Fig. 2, it was a “simpler” data set, in which the major difficulty was associated with the hierarchical structure.

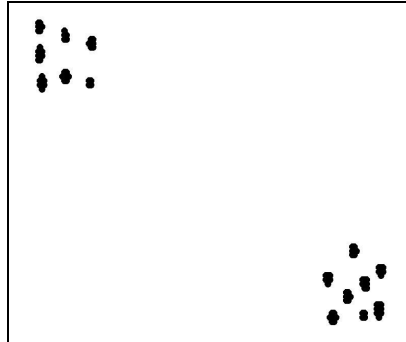


Fig. 2. A „simple” case treated by the algorithm proposed, with $n = 106$, $m = 2$

In this case the new algorithm was used with $p^1 = 45$ and, again, distances in D^1 calculated with option (2). The solution provided was that into two clusters, and the one into 15 clusters could only be identified via an additional analysis of the agglomeration diagram. In both cases, though, the clusters obtained were fully conform to eye inspection.

On the other hand, the classical k-means produced for $p^2 = 2$ the same result, in accordance with the image, and performed relatively well for $p^2 = 15$, although, for standard reasons, committed some misclassification errors.

4 Conclusions

So, a simple hybrid clustering scheme was implemented, which can be used effectively, as this was verified on a series of examples, for visualization purposes, and, in the same vein, for pattern recognition (also well beyond $m=2$). It can, of course, also be used as a general purpose clustering algorithm. In the examples treated to date the use of approximation in the calculation of distances D^1 for the second stage of the algorithm has not resulted in “incorrect” results.

We have on purpose not quoted here the computational statistics, because, first, these depend heavily on the choices made by the user (especially p^1) and can easily change the order of magnitudes, and second – we have been, at this stage of work, mainly interested in the uncovering of the “true” structure of data.

The work on the method is being continued. It concerns both the important technical details, commented upon in the paper, and some more general issues. These include, first of all, the final selection of p^2 on the basis of the agglomeration diagram (a classical problem, here its solution being assisted by application of the objective function from [3] and [4]), and the tuning of the second stage (e.g. a choice of the progressive merger procedure according to the Lance-Williams-Jambu formula) oriented at identification of various kinds of shapes of the clusters sought.

References

1. Marcotorchino F., Michaud P.: *Optimization in ordinal data analysis*, IBM France Scientific Centre. Technical Report, Paris (1978).
2. Mirkin B.: *Mathematical Classification and Clustering*. Kluwer Academic Publishers, Dordrecht Boston London.
3. Owsinski J. W.: *On a quasi-objective global clustering method*, In: Diday E., Jambu M., Lébart L., Pagés J., Tomassone R. (eds.): *Data Analysis and Informatics III*. North Holland, Amsterdam (1984) 293-306.
4. Owsinski J. W.: *On a new naturally indexed quick clustering method with a global objective function*, *Applied Stochastic Models and Data Analysis*, 6 (1990).