POLSKA AKADEMIA NAUK
Instytut Badań Systemowych

BADANIA OPERACYJNE I SYSTEMOWE: ŚRODOWISKO NATURALNE, PRZESTRZEŃ, OPTYMALIZACJA

Olgierd Hryniewicz,
Andrzej Straszak,
Jan Studziński
red.

BADANIA OPERACYJNE I SYSTEMOWE: ŚRODOWISKO NATURALNE, PRZESTRZEŃ, OPTYMALIZACJA

INSTYTUT BADAŃ SYSTEMOWYCH • POLSKA AKADEMIA NAUK

Seria: BADANIA SYSTEMOWE
tom 63

Redaktor naukowy:

Prof. dr hab. inż. Jakub Gutenbaum

Warszawa 2008

Olgierd Hryniewicz, Andrzej Straszak, Jan Studziński

BADANIA OPERACYJNE I SYSTEMOWE: ŚRODOWISKO NATURALNE, PRZESTRZEŃ, OPTYMALIZACJA

Publikacja była opiniowana do druku przez zespół recenzentów, którego skład podano w treści tomu

Opinie, wyrażone przez autorów w pracach, zawartych w niniejszym tomie, nie są oficjalnymi opiniami Instytutu Badań Systemowych PAN, ani Polskiego Towarzystwa Badań Operacyjnych i Systemowych.

ISBN 83-894-7519-7
EAN 9788389475190

Redakcja i opracowanie techniczne: Jan W. Owsiński, Aneta M. Pielak, Anna Gostyńska

Lista recenzentów
artykułów, wchodzących w skład tomów serii „Badania Systemowe”
związanych z konferencją BOS 2008

Dr Paweł Bartoszczuk
Dr inż. Lucyna Bogdan
Dr hab. inż. Zbigniew Buchalski
Mgr inż. Hanna Bury
Prof. dr hab. Marian Chudy
Dr Jan Gadomski
Mgr Grażyna Grabowska
Mgr inż. Andrzej Jakubowski
Dr hab. inż. Ignacy Kaliszewski
Dr Andrzej Kałuszko
Dr hab. Leszek Klukowski
Dr hab. inż. Wiesław Krajewski
Dr inż. Lech Kruś
Dr hab. inż. Marek Libura
Dr Barbara Mażbic-Kulma
Dr inż. Edward Michalewski
Dr inż. Jan W. Owsiński
Dr inż. Grażyna Petriczek
Dr inż. Henryk Potrzebowski
Dr Maciej Romaniuk
Prof. dr hab. Piotr Sienkiewicz
Dr hab. Henryk Spustek
Prof. dr hab. Andrzej Straszak
Dr hab. inż. Jan Studziński
Prof. dr hab. Tomasz Szapiro
Mgr Anna Szediw
Dr inż. Grażyna Szkatuła
Dr hab. inż. Tadeusz Witkowski
Dr Irena Woroniecka-Leciejewicz
Dr hab. Sławomir Zadrożny
Dr inż. Andrzej Ziółkowski

Komitety Konferencji
Badania Operacyjne i Systemowe 2008
Rembertów, Akademia Obrony Narodowej

Patronat honorowy

Bogdan Klich, Minister Obrony Narodowej
Maciej Nowicki, Minister Środowiska i Zasobów Naturalnych

Komitet Sterujący

Janusz Kacprzyk, Prezes Polskiego Towarzystwa Badań Operacyjnych i Systemowych
Olgierd Hryniewicz, Dyrektor Instytutu Badań Systemowych
Janusz Kręcikij, Komendant Akademii Obrony Narodowej

Komitet Programowy

Piotr Sienkiewicz, *Przewodniczący*
Jacek Mercik, *Wiceprzewodniczący*

Tomasz Ambroziak
Marian Chudy
Olgierd Hryniewicz
Ignacy Kaliszewski
Piotr Kulczycki
Barbara Mażbic-Kulma
Włodzimierz Ogryczak
Andrzej Piegat
Honorata Sosnowska
Andrzej Straszak
Ryszard Tadeusiewicz
Jan Węglarz
Ryszard Budziński
Ludosław Drelichowski
Adam A. Janiak
Józef Korbicz
Małgorzata Łatuszyńska
Zbigniew Nahorski
Wojciech Olejniczak
Krzysztof Santarek
Henryk Spustek
Tomasz Szapiro
Eugeniusz Toczyłowski
Tadeusz Witkowski
Bogdan Zdrodowski
Wojciech Cellary
Jerzy Hołubiec
Jerzy Józefczyk
Maciej Krawczak
Marek J. Malarski
Andrzej Najgebauer
Jan W. Owsiński
Roman Słowiński
Jan Stachowicz
Andrzej Szymonik
Tadeusz Trzaskalik
Stanisław Zajas

Komitet Organizacyjny

Jan W. Owsiński, Andrzej Kałuszko, Mieczysław Pelc, Zbigniew Piątek

Sekretariat

Krystyna Warzywoda, Monika Majkut, Aneta M. Pielak, Krzysztof Sęp, Anna Stachowiak, Halina Świeboda, Tadeusz Winiarski

Redakcja wydawnictw

Janusz Kacprzyk, Piotr Sienkiewicz, Andrzej Najgebauer, Olgierd Hryniewicz, Andrzej Straszak, Jan Studziński, Jan W. Owsiński, Zbigniew Nahorski, Tomasz Szapiro

Metody:
Optymalizacja, dane, analiza

A NEW HYBRID CLUSTERING METHOD: FROM "SUBASSEMBLIES" TO "SHAPES"

Jan W. Owsiński[1] **and Mariusz Tomasz Mejza**[2]
[1] Systems Research Institute, Polish Academy of Sciences, Newelska 6,
01 447 Warsaw, Poland
owsinski@ibspan.waw.pl
[2] National Bank of Poland, Warsaw, Poland

The paper presents a simple hybrid technique of clustering, based on the consecutive use of k-means and agglomerative (e.g. nearest neighbour) algorithms. The technique starts with production of "subassemblies" through application of the k-means algorithm, performed as the first stage for an adequately high number of centroids, and continues with identification of "shapes", by using an agglomerative (say, nearest neighbour) algorithm, executed for the clusters obtained in the first stage, as the set of initial objects to be merged. A simple analysis of computational complexity shows the differences among variants of the new technique, depending upon the manner, in which distances among the "subassemblies" are calculated. It is shown on examples how the thus defined technique performs in terms of identification of relatively complex shapes.

1 Introduction

We shall present in this paper a new clustering technique, designed so as to overcome some of the shortcomings of the techniques existing as of now, while retaining their known strong points. It is a relatively simple hybrid technique, associating the principles of the k-means and agglomerative schemes. In the first stage of the technique, k-means-type algorithm serves to produce the "subassemblies" of the potential "shapes", to be identified in the second phase, as proposed here – performed through the use of an agglomerative scheme. Besides presenting the technique and its effectiveness for a relatively broad class of clustering problems, the paper takes up some issues of computational complexity with respect to the technique.

1.1 The clustering problem(s)

Clustering, i.e. *placing the similar together and the dissimilar apart*, is not only a model of the basic intellectual activity, and not only one of the fundamental problems 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 Marcotorchino and Michaud (1978) or Owsiński (1984, 1990), 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, see,

e.g., Rządca (2004). 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 Mirkin (1996, 2005).

The dozens of existing approaches have each some merits in one or more of the application fields or the aspects of quality (e.g. interpretation of results, identification of patterns, [extreme] computational simplicity meant for the very large data sets, etc.), but also display, inevitably, poor performance with respect to some of the 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 purposes of developing a new technique

The reasons for the development of the new technique can be summarized as follows: (1) to develop a method of clustering that provides 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 intervened in the sequence as here provided: from a more modest goal to a much more ambitious one. It was, namely, hoped that effective visualization, without any undue distortion, 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 new technique

We shall now give the complete description of the approach, with a short consideration of the technical details, primarily those related to computational effectiveness and efficiency.

2.1 Notation

Assume we deal with n objects (observations, items), indexed i, $i \in I = \{1,\ldots,n\}$. 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, corresponding to the set of objects considered, is divided (partitioned) in the clustering problem into subsets (clusters) denoted A_q, $q = 1,\ldots,p$, where p, or $p(P)$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 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 $p_1 = n^{1/2}$. So, clusters $A^1{}_q$ are obtained, $q = 1,\ldots,p_1$.

Once the first stage terminated, the matrix of distances between clusters $A^1{}_q$ is calculated, D^1. On the basis of this matrix a classical progressive merger procedure is performed, in the very first implementation of the algorithm – the single link (nearest neighbour) procedure.

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 afterwards on the basis of some additional information.

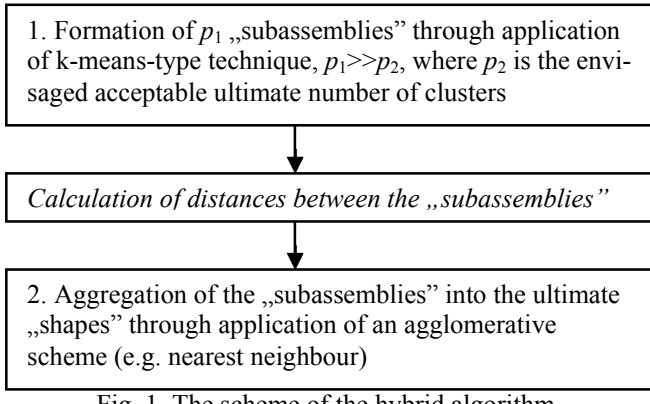

Fig. 1. The scheme of the hybrid algorithm

2.3 The reasons behind

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 forms the spherical or ellipsoidal clusters, is computationally efficient and converges quickly. If the clusters formed are "small", and "dense" in the set of objects considered, the convergence is (relatively) 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. Thereby, the "subassemblies" are formed of the potential ultimate "shapes" to be recovered. Since it is assumed that they are just elements of the ultimate clusters, it is not so important to fine-tune the first stage of the algorithm, spending additional time and memory on such fine-tuning (see Section 3).

In the second stage an agglomerative scheme, here the single link, is used, which aggregates the subsets according to minimum distances, so that if these subsets form any linear and complex shape, it should get uncovered, with the agglomerative algorithm not so much penalised by the necessity of maintaining and recalculating the distance matrix, owing to the shrinking of the dimension of the problem in the first stage.

2.4 The technical issues

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, which is meant to secure equal proportions between n, p_1 and the final outcome of the classical agglomerative scheme, i.e. p_2=1, caution must in general be made of the maintenance of 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 (at least by an order of magnitude), the initial centroid candidates can be determined by a method different from the usual random choice in E_X (or X_I), and different from the initialisation methods, based on density assessment, used in some implementations of k-means;
- **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 several 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, and
(**4**), the least requiring, the distance used is $d(A^1_q,A^1_{q'}) = d(x^q,x^{q'})$, i.e. simply, the distance between the centroids.

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, in terms of both time and memory requirements (see next section). Beyond purely computational aspect, though, this issue has also a more "profound" significance: namely, it can be proposed that the definitions of D^1 correspond to the subsequently performed agglomerative scheme. And so, in particular, the definition (1) above corresponds to the case of single linkage, while if definition $d(A^1_q,A^1_{q'}) = \max \{d_{ij}: x_i \in A^1_q, x_j \in A^1_{q'}\}$ were adopted, it would correspond to the case of complete linkage. In this manner it is possible to design the definitions of D^1 corresponding to the subsequent agglomerative schemes. Such a "matching", though, was not considered necessary at this stage of work, since the clusters A^1_q result already from a different kind of scheme and there is no "theoretical" prerequisite for ensuring such a "matching".

3 Computational efficiency

The subsequent section shows the examples, for which the algorithm produced the "precise" results expected, which are hardly obtainable by most – if not all – of the existing clustering algorithms. While admitting, definitely, that these results are impressive, the question may be asked of the (additional) computational burden, necessary to produce them. In the following we shall focus primarily on the number of distance calculations, as the most complex of the "unit" operations in the algorithms of clustering.

Thus, in the first stage the order of the thus conceived complexity is $O(np_1)$, while in the second – $O(p_1^2 \log p_1)$. Hence, for the hint of $p_1=n^{1/2}$, we have $O(n^{3/2})$ and $O((n\log n)/2)$. Lowering of p_1 naturally leads to lower complexity of these two stages, getting closer to that of the k-means techniques. The stage of calculating distances D^1 strongly depends upon the choice of options (1) through (4). In case of option (1) the complexity is $O(p_1(n/p_1)^2) = O(n^2/p_1)$, that is – when $p_1=n^{1/2}$, $O(n^{3/2})$. In the simplest of options, (4), this complexity is $O(p_1^2)$, i.e. for the hint, $O(n)$. Hence, we deal with the range of complexities between $O(np_1+n^2/p_1+p_1^2\log p_1)$ and $O(np_1+p_1^2(1+\log p_1))$, and for $p_1=n^{1/2}$, between $O(2n^{3/2} + (n\log n)/2)$ and $O(n^{3/2}+n + (n\log n)/2)$, which means that the ranges are not very broad, but definitely, p_1 is an essential decision variable with this respect, in connection with the choice of distance calculation option. Thus, by selecting p_1 and the distance calculation option, we may essentially influence computational complexity involved, and the additional

burden imposed is not really significant, once we accept the initial conditions, resulting from the use of k-means and agglomerative schemes.

Let us add at this point that we used the standard procedures in both of the essential stages of the algorithm. There certainly exist refinements or extensions to, for instance, k-means, which either reduce the computational burden (see, e.g. Daschiel, Datcu, 2003), or attempt to grasp more elaborate cluster shapes than spherical or ellipsoidal (e.g. Dhillon, Guan, Kulis, 2005), but the cost of identifying these more elaborate – usually largely pre-defined – shapes is tangible (complexity rising beyond $O(n^2)$), while the advantages of the simplified k-means techniques are, in general case, quite limited.

4 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. 2. A „clinical" case treated by the algorithm proposed, with $n = 2277$, $m = 2$

For the example shown schematically in Fig. 2 above, with relatively complex shapes of the clusters "to be identified", the new algorithm, with the choices of p_1 ranging from 60 to 227 and distances in D^1 calculated according to option (2), allowed, in all cases, for the precise identification of the visually obvious clusters on the basis of the aggregation distance diagram. The classical k-means, for which the "correct" value of $p^2 = 6$ was set, was, of course, 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. 3, it was, in a way, a "simpler" data set, in which the major difficulty was associated with the hierarchical structure.

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.

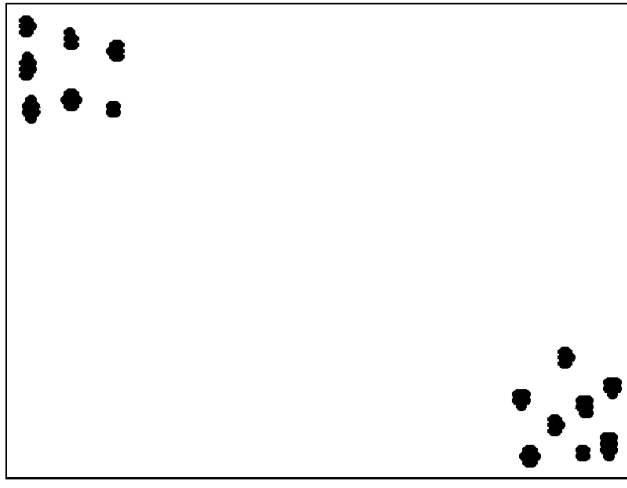

Fig. 3. A „simple" case treated by the algorithm proposed, with $n = 106$, $m = 2$

5 Conclusions and future work

A general hybrid clustering scheme was proposed and then implemented through the use of standard techniques of the respective component algorithms, k-means-like and agglomerative. This scheme can be used effectively, as 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.

The scheme is both general and flexible. It can accommodate different variants of the basic algorithms used in the scheme, and different methods of calculating distances between the "subassemblies".

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, the final selection of p^2 on the basis of the agglomeration diagram (a classical problem, its solution being assisted in the future work by application of the objective function from Owsiński, 1984, 1990), the various possibilities of calculating distances D^1 (or similarities S^1) between p_1 clusters, and the more elaborate 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

Daschiel, H., Datcu, M. P. (2003) 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, vol. 4885, 120-130.

Dhillon, I., Guan, Y., Kulis, B. (2005) A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts. UTCS Technical Report #TR-04-25, University of Texas, Austin, TX.

Marcotorchino, F., Michaud, P.: Optimization in ordinal data analysis. IBM France Scientific Centre. Technical Report, Paris (1978)

Mirkin, B. (1996) *Mathematical Classification and Clustering*. Kluwer Academic Publishers, Dordrecht Boston London.

Mirkin, B. (2005) *Clustering for Data Mining. A Data Recovery Approach*. Chapman & Hall, London.

Owsiński, J.W. (1984) 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, 293-306.

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

Rządca, K. (2004) Algorytmy grupowania danych (Algorithms of data grouping; in Polish). M.Sc. thesis. Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Warszawa.

Olgierd Hryniewicz, Andrzej Straszak, Jan Studziński
red.

BADANIA OPERACYJNE I SYSTEMOWE: ŚRODOWISKO NATURALNE, PRZESTRZEŃ, OPTYMALIZACJA

Książka składa się z artykułów przedstawiających wyniki prac z dziedziny badań operacyjnych i systemowych, poświęconych środowisku naturalnemu i zarządzaniu nim, zwłaszcza w zakresie ochrony atmosfery, globalnego ocieplenia i walki z nim, jakości i zaopatrzenia w wodę. Tematyka ta jest rozszerzona o aspekty przestrzenne, regionalne i samorządowe, a także planowanie i funkcjonowanie infrastruktury. Tom zamykają prace metodyczne, dostarczające technik, będących podstawą prezentowanych zastosowań.

ISBN 83-894-7519-7
EAN 9788389475190

Instytut Badań Systemowych PAN
tel. (4822) 3810241 / 3810273 e-mail: biblioteka@ibspan.waw.pl