DEVELOPMENT AND REALIZATION OF AGGLOMERATIVE REGIONALIZATION ALGORITHMS AND PROBING OF THEIR TIME COMPLEXITY

Vladimir MESURA
Vinnica Technical State Univercity
vimes@vstu.vinnica.ua
Andrew ZAVARZIN
GIS-Association, Moscow
gisa@inbox.ru

Abstract. *There are a number of mentions in the science literature about the usage of agglomerative hierarchical algorithms for regionalization. Usually, described algorithms are based on one of the distances between regions in a factor space (for example, group average).*
The given article contains generalization of classical classification agglomerative hierarchical algorithms for task of regionalization. In the given article, modes of their realization and the results of experiments under the analysis of their time complexity are described.
Keywords: *regionalization, spatial data, agglomerative hierarchical algorithms.*

Introduction to the regionalization task

Multivariate spatial data is a number of objects which are located simultaneously both in geographical and in factor spaces. The factor space is usually represented by the table object-factor. This table depicts the measurement of M factors on N objects and containing N lines and M columns. The geographical space is usually represented by a layer of digital map in GIS. In geography, the purpose of classifications is to obtain of some number of objects groups $S_1,...,S_K$ by these information. In each of class objects should be similar, homogeneous against each other. Objects from different classes must be maximum various.

Frequently at the analysis of spatial data it is required to solve the task of regionalization. Regionalization is the task of territory division into a set of the not intersected entire regions representing compact objects concentration both in geographical, and in factor spaces [1, page 3]. The result of regionalization can be represented by N-dimensional vector $v = (v_1,..., v_N),\ v_i \in \{1,..., K\}$:

$$v_i = j \Leftrightarrow o_i \in S_j,\ i \in \{1,..., N\},\ j \in \{1,..., K\}.$$

Many of agglomerative algorithms of regionalization are based on appropriate methods of classification. The difference is in adding the procedure of geographical adjacency check of joined classes. Sequence of partitions of a set of objects O on not intersected classes $\{S^n \mid n = 1,..., I+1\}$ is result of hierarchical classification algorithms working, where

I - the number of spent iterations,

S^n - one of the system of classes.

The main difference between the number of hierarchical agglomerative classification algorithms is in a mode of calculation of distances between classes. Classes distance D sometimes are called strategies of classes join. They are always based on distance d between single objects from two classes. They also may be determined by various ways.

Most usefull distances D:

1. Method of the nearest neighbour $D_{\min}$. The distance between two classes calculates as the distance between two the nearest objects of these two classes.
2. Method of the long-distance neighbour $D_{\max}$. The distance between two classes

calculates as the distance between two long-distance objects of these classes.

3. Centroid method D_{cen}. The distance between two classes calculates as the distance between to classes centers.
4. Group average method D_{avg}. The distance between two classes calculates as the average distance between all objects of two classes.

The calculation speed of each described distances is important for the further astimate of algorithms time complexity. Let N_i, N_j - the number of objects in each of two classes. The order of time complexity depends on how many distances it is necessary to calculate for intergroup distance receiving.

Table 1. The order of time complexity of intergroup distances calculation.

D	**Time complexity**
$D_{\min}$	$O(N_i N_j)$
$D_{\max}$	$O(N_i N_j)$
D_{cen}	$O(N_i)+O(N_j)$
D_{avg}	$O(N_i N_j)$

The scheme of suggest regionalization agglomerative algorithms

Algorithms schemes with an existing mode of an intergroup distances evaluation are reflected in many research works dealing with the technique of regionalization. It is expedient to generalize classical algorithms of classification for the purposes of regionalization.

Having defined above possible modes of intergroup distances calculation, we'll show agglomerative algorithm updated for the purpose of regionalization.

1. To choose metric d and intergroup distance D.
2. To form the first system S^1 which is consist of N regions:
   $S^1 = \{S_1^1, \ldots, S_N^1\}$, $S_1^1 = \{o_1\}$, ..., $S_N^1 = \{o_N\}$.
   Set $n = 1$.
3. To set $K = N - n + 1$ and calculate contiguity matrix $G_{K\times K}^n$:
$$g^n(i,j) = \begin{cases} 1, \exists o_{ix} \in S_i^n, \exists o_{jy} \in S_j^n : g(o_{ix}, o_{jy}) = 1; \\ 0, \; else \end{cases}$$
   $i, \; j \in \{1, \ldots, K\}$
4. Suppose that on the step $n \in \{1, \ldots, N-1\}$ we have the system of regions
   $S^n = \{S_1^n, \ldots, S_K^n\}$.
   Then:
   4.1) To calculate $D(S_i^n, S_j^n) \; \forall i, j \in \{1, \ldots, K\}$.
   4.2) To find $x \neq y, \; x, y \in \{1, \ldots, K\}$:
$$D(S_x^n, S_y^n) = min\{D(S_i^n, S_j^n) \mid i \neq j, \; g^n(i,j) = 1\}$$
   4.3) Let $i < j$. Set
$$S_i^{n+1} = \begin{cases} S_i^n, \; i \in \{1, \ldots, y-1\} \setminus \{x\}, \\ S_x^n \bigcup S_y^n, \; i = x, \\ S_{i-1}^n, \; i \in \{y+1, \ldots, K\} \end{cases}$$
5. If $n + 1 = N$ then stop. Else set $n = n + 1$ and go to step 3.

For research into the usage of expediency of the algorithms class described by the authors, algorithms were realized in the software product of multivariate space data classification called GisCluster. After that the number of experiments were conducted. The purpose of the experiments were:

- automatic regionalization and expert rating of regionalization quality;
- comparison of regionalization algorithms for mining of their specific properties;
- research on algorithms time complexity.

Just point out, that all of the algorithms allow either to expose some features in data or gave good final regions schemes.

Realization variants of the algorithms

The most time-consuming stage of regionalization algorithm is recalculation of distances between regions on each step n:
$D(S_i^n, S_j^n) \; \forall i, j \in \{1, \ldots, K\}$.

At realization of the algorithms two alternate approaches were considered. The first parsed approach is based on usage of Jambue formula. This formula allows to calculate new distances between classes on a basis of already calculated [2, page 497]:

$$D(S_i,\ S_x \cup S_y) = \alpha D(S_i, S_x) + \beta D(S_i, S_y) +$$
$$+ \gamma D(S_x, S_y) + \delta \left| D(S_i, S_x) - D(S_i, S_y) \right| \quad (1)$$

Index of iteration n is not shown in the formula (1). At various values of parameters $\alpha, \beta, \gamma, \delta$ this formula corresponds to an evaluation useful intergroup distances. Values of coefficients are known for each intergroup distance.

Realization of classical classification agglomerative algorithms on the basis of the formula (1) greatly increases their speed. However, for its usage in the regionalization algorithm it is necessary to calculate all intergroup distances on each step. Thus, it is possible to think that there will be a significant loss in an operating time on the first steps of algorithm.

To overcome the indicated weakness, at realization of GisCluster software another approach was used. It reduces an amount of intergroup distances evaluations.

1. The constant D_∞ is defined. It is the maximum positive value by which non-adjacent regions in a matrix of intergroup distances are marked. To the regions which don't have common boundary, certainly inaccessible maximum distances are given $D_\infty = \{D(S_a, S_b) \mid g^n(a,b) = 0\}$.
2. On the next step, after finding a minimum in a matrix of distances between two regions S_x and S_y and their joining in one region $S_x \cup S_y$, only distances recalculations are made: $D(S_x \cup S_y, S_z)$, where $g^n(xy, z) = 1$.
3. The last condition corresponds either to the case $g^n(x,z) = 1$ ($D(S_x, S_z) < D_\infty$), or to the case $g^n(y,z) = 1$ ($D(S_y, S_z) < D_\infty$).

Therefore at recalculation of distances in the formula (1) some addends are known. Recalculation of the rest distances of the formula (1) is carried out with the usage of M-dimensional vectors of objects of each classes. It is one of the most toilful operations and lack of the second approach.

As it follows from exposition, the amount of evaluations of intergroup distances is reduced. For example, on a first step from $N-2$ up to $\Theta(S_{xy})$. $\Theta(S_{xy})$ is the number of regions, adjacent with the given region S_{xy}. Usually $\Theta(S_{xy}) \le 15$, but very often $\Theta(S_{xy}) = 6$.

Decribtion of the experiments of algorithms time complexity

The authors fulfilled a number of the experiments for approbation of realized algorithms on the real data for the estimation of their time complexity. The purposes of experiments were exposition of association of the algorithm operating time from structures of factor and geographical spaces, and also from parameters D, M, N and n. The operating time of the algorithms realized on the basis of Jambue formula with total recalculation of all distances on each step should not depend on structure of factor and geographical spaces, D and M.

We used the statistics on federal elections of the Russian Federation. There were data about subjects of the Russian Federation (89 objects) and the elective commissions (about 2500 objects).

Calculations were carried out by the personal computer with the processor such as PIII-500 and the RAM of 128 megabytes. The some results for time function $T(n)$ for different count of objects $N \in \{89, 2500\}$ and $M = 5$ are show below.

Table 2: Some samplings of the realized algorithms operating time.

№	N	n	D	T, sec.
1	89	89	$D_{\min}$	1
2	89	89	$D_{\max}$	1
3	89	89	D_{cen}	1
4	89	89	D_{avg}	1
5	2500	2300	$D_{\min}$	4504
6	2500	2300	$D_{\max}$	451
7	2500	2300	D_{cen}	371
8	2500	2300	D_{avg}	1361

It is clear from table 2, at small sizes of data (89 objects) the algorithms operation time is quite acceptable. The user receives results instantly. For many objects table 2 shows only the grade of time complexity. It is necessary to determine:

- the type of dependence of the time function $T(n)$ for each algorithm;
- behaviour $T(n)$ on different data for the same dimension;
- behaviour $T(n)$ depending on a various amount of factors M.

All further calculations are for the case of 2500 objects.

Comparison of an algorithms operating time among themselves

As follow from the table 2 it is possible to make performance about complexity of various methods.

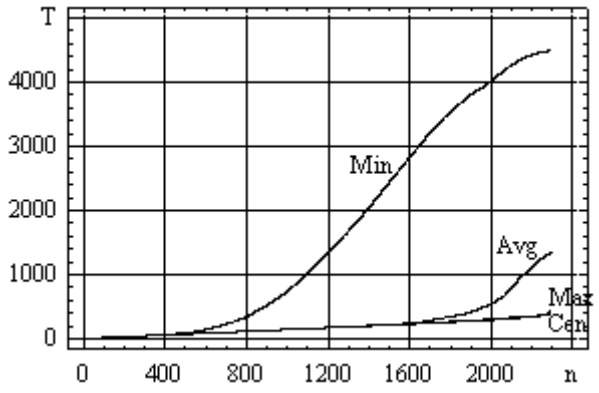

Figure 1. Behaviour of the function $T(n)$ for various kinds of intergroup distances D.

By the results of experiments it is possible to suppose, that algorithm time complexity depends on complexity of an evaluation of intergroup distances and depends on symmetry of obtained regions. At the initial stages of operation ($n \in [0, 600]$) all algorithms give approximately identical time results.

Sharp magnification of an operating time of algorithm for $D = D_{\min}$ follows, obviously, from specificity of the given algorithm. This specificity often appears during creation one big representative region with many objects in it. At visual review the course of nearest neighbour algorithm is similar to creation snow ball.

Algorithm with $D = D_{\max}$ often creates symmetry regions. Therefore time of its operation is the same as in centroid method. The operating time of any algorithm also depends on the structure of geographical space. The more contiguities it contains, the more complicated and longer calculations are.

Analysis of different data with the same dimension

For behaviour $T(n)$ research for each type of distance D At $M = const$ the number of experiments were conducted. They were generalized as follow.

Table 3. Sampling of an operating time $T(n)$ for $D = D_{cen}$ and different factors at $M = const = 5$.

n	1	…	500	…	1000	…	2300
T_1^{cen}	1	…	72	…	144	…	408
…	…	…	…	…	…	…	…
T_J^{cen}	1	…	72		143		371

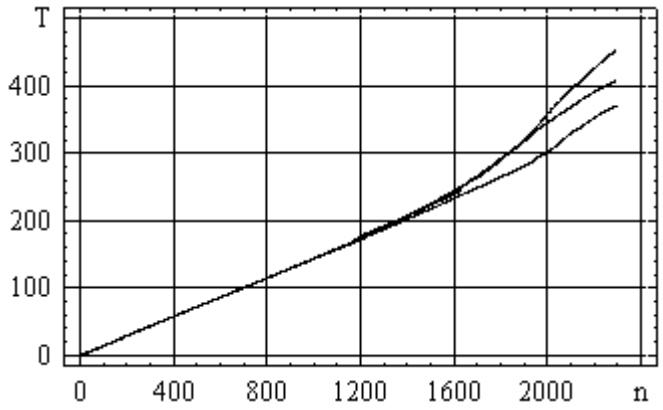

Figure 2. The graph of an operating time $T(n)$ for $D = D_{cen}$ and three implementations on different factors at $M = const = 5$.

The results of experiments shown, that the operating time of all algorithms varies in some reasonable limits at passage from one set of factors to another at $M = const$. It is becouse of the various configuration of formed regions in geographical and factor spaces at each implementation.

It is interesting to research an operating time of the same algorithm for the same data, but for different geographical spaces. However, such experiment is practically impossible to carry out.

Dependence $T(n)$ on amount of M factors

Having a number of experiments for the same algorithms on the data with different count of factors $M \in \{1,\ldots,15\}$, the existence of some positive dependence between T And M was disclosed.

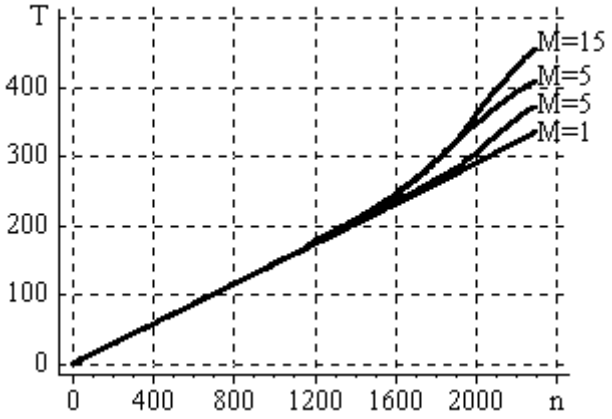

Figure 3. Dependence $T(n)$ for $D = D_{cen}$ and $M \in \{1,5,5,15\}$

Experiments were not carried out for dimension $M > 15$. If $M > 15$, it is necessary to use methods of factor and principal component analysis.

Conclusions

The developed hierarchical agglomerative methods of regionalization are briefly described in the article. It is described two modes of algorithms realizations: on the basis of Jambue formula with total recalculation of distances and on the basis of calculation of interdistrict distances only for geographically adjacent objects. Some of the experiments are also described, with indicating the growth of time depending on data size. It is indicated, that even at handling thousands objects this time is accessible (some minutes). Interesting results are obtained, for example, at comparison an operating time of the nearest and long-distance neighbour algorithms on the same data. At absolutely identical realization, the long-distance neighbour algorithm works almost twice faster. It is becouse of its properties on creation symmetry regions.

Perspective direction of research is to comparison of an operating time of the realized algorithms with algorithms on the basis of Jambue formula. It is interesting to detect the number of factors M, number of objects N, average of geographical contiguities $E(\Theta)$,

when this or that realization of algorithm should be used.

Acknowledgement

Authors express their gratitude to the professor of the Moscow State University doctor V.Tikunov for long-term support. We are also appreciated D.Oreshkina, who supplied data for the analysis.

References

[1] Blanutsa, V.I. (1993) Integral ecological regionalization: the concept and methods. - Novosibirsk, "Science".
[2] Aivazyan S.A, Mhitaryan V.S. (2001) Foundations of econometrics. Probability theory and applied statistics. - M., Unity, 656 pages.
[3] Tikunov V.S. Simulation in cartography. (1997) - M.: Moscow State University, 405 pages.