Method Of Abaft Proliferation – A Fuzzy Based Region Segmentation Technique

Anirban Chakraborty

Assistant Professor
Department of Computer Science,
Barrackpore Rastraguru Surendranath College, Barrackpore, Kolkata -120, W.B., India

Abstract: In this paper a methodology for region segmentation been proposed. It takes as input a digitized map and the number of segments to form. It calculates the association of each constituent region with the centroids determined initially. A concept of fuzziness is associated here. Judging both fuzziness and the associated values of population, segmentation is formed so that each has approximately same population. Care been taken to remove the virtual clusters also.

Index Terms - **Region segmentation, fuzziness, associated value, digitized map, virtual cluster.**

I. INTRODUCTION

Region segmentation plays a crucial role in GIS [4, 5, 6]. Segmentation of regions (generally done on input maps) is done to accommodate similar type of regions within one segment. If a map with many constituent regions are considered, then after segmentation, each segment will hold similar type of regions. Concept of map partitioning is almost same as region clustering. In most of the GIS anchored partitioning techniques, user targets to produce partitions in such a manner, so that either each partition will be of same area or each will hold same amount of attribute data, such as population.

The 'Method of Abaft Proliferation' initially takes as input a raster map, constituted by a number of of regions, population of each of which is known and the objective is to divide the input raster map into a number of segments, such that each segment contains approximately equal amount of population, or in other words the entire population has to be divided among the segments. Moreover, each segment should be composed of adjacent regions, i.e. regions which share some common geographical line. As its basic underlying concept, this segmentation technique uses the concept of fuzziness, which increases the compactness of the segments formed. This proposed method actually works in two distinct phases. In its first phase, based on the distances between region centroids, it performs segmentation, without taking into account of associated attribute data, to increase compactness of segmentation. In the second phase, the size of the segments are adjusted to distribute the population among segments in a uniform way.

The first phase of the proposed segmentation procedure is driven by considering the membership value of the regions to various segments, being calculated on the basis of point to point distance between the segment's centroid and region's centroid. More a region (i.e. its centroid) is closer to a segment (i.e. center of the segment); its membership value towards that particular segment is also larger. A Membership Matrix is used for the purpose. If N is the total number of regions in the map considered and k segments are to form, then the Membership Matrix contains k rows and N columns, where Cell (i,j) holds the membership value of the jth region towards the ith segment.

II. METHODOLOGY

Initially as segment centroids, k region centroids are arbitrarily chosen . Here, if the border region's centroid (far apart from each other, for example left-most, top-most etc.) are chosen, then less number of iterations are needed for convergence of the process. The Membership Matrix is also populated with some random data. At each stage the centroid to- centroid distance between segment's centroid and region's centroid is calculated. On basis of these distance values, the Membership Matrix, along with the Segment Centroids are updated using the following equations.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{\left(\frac{2}{m}-1\right)}}$$

$$v_j = \frac{\sum_{i=1}^{n} (\mu_{ij})^m x_i}{\sum_{j=1}^{n} (\mu_{ij})^m}, \forall j = 1,2,....,c$$

Here, n is the number of points (region centroids) dealt by equations shown above. The segment centroid of jth segment is denoted by v_j, parameter m can take any value between 1 to ∞. Presently m = 2 been chosen. The number of segments to be formed is denoted by c. The association between ith region and jth segment is denoted by μ_{ij} and finally the distance between i^{th} region and centroid of jth segment is represented by d_{ij}. Similarly, the distance between i^{th} region and k^{th} cluster centroid is represented by d_{ik}.
Actually here the main objective is to minimize:

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m \|x_i - v_j\|^2$$

where, n is n is the number of regions, c is the number of segments to be formed, U = [μ_{ij}], V is the group of centres of c segments, x_i is the ith region (i.e. its centroid) and v_j is the jth cluster centroid.

Let $Cell_t(i, j)$ and $Cell_{t+1}(i, j)$ denotes the value of Cell (i,j) in the Membership Matrix at t^{th} and $(t+1)^{th}$ iterations respectively. When for all cells of the membership matrix ($|Cell_{t+1}(i, j) - Cell_t(i, j)|$) ≤ β (where β is the predefined value of stopping criterion); the first phase of the segmentation procedure stops. At this stage, the Membership Matrix is observed column wise. For any column j, if the maximum value exists at row i, then j^{th} region becomes the part of ith segment. If this maximum value exists for more than one row, any one among them could be arbitrarily chosen. The concept has been explained as follows.

Let five regions are to be divided among two segments. At the stage, when the first phase of the method terminates, let the Membership Matrix contains the values shown in table 1.

Table 1: Contents of the Membership Matrix

Segments	Regions				
	1	2	3	4	5
Segment 1	0.4	0.8	0.9	0.4	0.7
Segment 2	0.6	0.2	0.1	0.6	0.3

These values directs to place Region 2, 3 and 5 at Segment 1 and Region 1 and 4 at Segment 2. In some situations, some rows at this steps may also found, for which no columns possess maximum values. Such a situation has been shown in table 2, where three segments are to achieve from five regions.

Table 2: Contents of the Membership Matrix (Exception condition)

Segments	Regions				
	1	2	3	4	5
Segment 1	0.3	0.7	0.7	0.3	0.6
Segment 2	0.6	0.2	0.1	0.6	0.3
Segment 3	0.1	0.1	0.2	0.1	0.1

It is quite obvious from table 2 that, no column is holding its maximum value at row 3. In other words, no region could be placed as a member of Segment 3. Thus although three segments were targeted, but practically two are obtaining, as the third one (Segment 3) is holding no members. In a general sense, while targeting for "k" segments, the above procedure may results a situation, where only "m" segments (m < k) are associated with data objects (regions) and rest "(k-m)" segments are virtual (in practical sense, as no data objects are associated with them). In such a stage, if any virtual segment(s) is generated, it is being removed incorporating the following procedure. Among the 'm' produced real-segments, the segment with maximum size (in terms of the number of constituent regions) is selected and simply bisected. Here 'bi-section' means the original segment S, containing 't' regions, is divided into 2 segments S_1 and S_2. Among which S_1 holds first ($\frac{t}{2}$) adjacent regions of S and S_2 holds rest ($\frac{t}{2}$) regions. (For an odd value of t, S1 holds 1 more region than S2). This results the number of real clusters to be 'm+1'. Next among the 'm+1' real segments, again the maximum sized segment is bisected. This process of bisection continues until 'k' real segments are achieved as per requirement. Here ends the first phase of the procedure, where as per user input, k real segments are produced. But in the entire tenure of the first phase, the segmentation is done only on basis of distance (to make each segment compact), without considering the value of associated attribute data (population) anyhow. The second phase of the procedure executes on the basis of associated attribute data (population), targeting an equal distribution of population among the segments. If N is the entire population considered (population of the whole map), then each of k segments should be of population $\frac{N}{k}$. However, as it is not permitted to divide any input region furthermore, so its not always possible to reach the exact value $\frac{N}{k}$ and that's why a deviation factor d comes onto the crease. A segment is announced to be complete, if its population lies into the range $\frac{N}{k} \pm d$.

If each of the k segments produced at the end of first phase, holds required amount of associated attribute data (population) then the procedure is over (which is practically a rarest situation). Otherwise among the k segments formed, the largest one (in terms of population) is chosen and if its population is more than $\frac{N}{k} \pm d$ then starting from the smallest populated, one by one of its constituent regions are eliminated until its population lies in the proper range (i.e. $\frac{N}{k} \pm d$). The eliminated regions are placed to the adjacent segments with least population, provided they are not complete. Once a segment is announced to be complete, it should not be changed anyhow. Thus, actually progressing backward, one after another segments are forcefully made complete. Hence the name "Abaft Proliferation" has used. During elimination of regions, care should be taken to exempt those regions from elimination, which could not be placed to any other segment as shown in figure1, where different segments are presented with different colors, in which regions marked with '×' are exempted from elimination. In other words, regions adjacent to other segments are only eliminated.
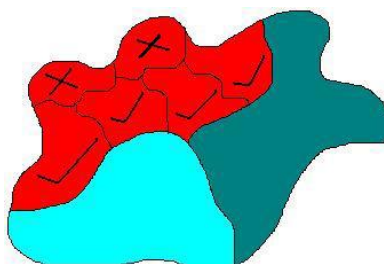

Fig. 1: Regions permissible for elimination

At any stage, when a segment is announced to be complete, during the entire rest part of the procedure, this is not changed anyhow. In other words, eliminated regions are placed only in incomplete segments, not being included in a complete segment. Thus if the first 'k-1'segments become complete, the rest one automatically completes, resulting 'k' complete segments as a whole. The procedure incorporated, could be summarized as state in Algorithm 1.

Algorithm 1 : Method of Abaft Proliferation

1: Assumptions:
  • Let R= {r1, r2, r3 . . . . . . rn} be the set of regions, to be grouped among k (k 6 n >2) segments
  • X = {x1,x2,x3 . . . . . .xk} be the set of data points, which are actually the centroids of the regions
  • V = {v1,v2,v3 . . . . . .vc} be the set of segment centers, initially chosen randomly
  • An Association matrix U reflects the association of a region with a segment. More is the association between a region with a segment, better is to place it there. Initially the matrix is filled by random values

2: The computation of membership μ_{ij} and the centers v_j are done at each iteration until the termination criterion $\| U(k+1) - Uk \| < \beta$ is reached

3: Inspecting the association matrix U; k segments are formed and among them some may be virtual, containing no regions.

4: Virtual Segments are removed by bisecting larger segments.

5: Starting from the populated most, one by one segment is inspected and by necessary elimination of the constituent regions from the segments, the population of each segment is made to lie within the required range, to make a uniform distribution of population among the segments.

Let 'n' is the number of constituent regions and 'c' is the number of segments to form. If the process converges after 'I' iterations then the time-complexity for the first phase is $\approx O(nc^2 I)$. For large data sets, where $c \ll n$, the time-complexity becomes $\approx O(n)$. Thus inclusion of the second phase makes the overall time-complexity $\approx O(n2)$.

III. RESULTS AND ILLUSTRATIONS

A step wise approach of implementation of the methodology in digitized map is explained illustrated using figure 2.
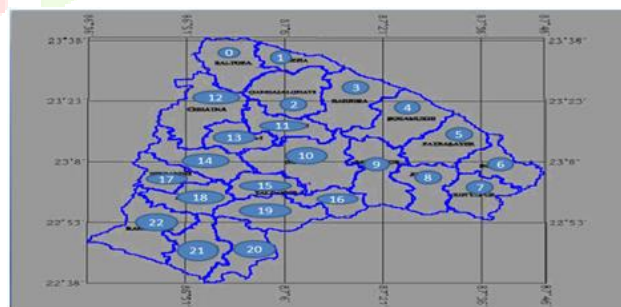

Fig. 2: Internally generated unique ID of the regions

Let us consider a digitized map of a very large area (Fig. 2), composed of 23 regions, identified by Region ids 0 to 22. These region ids are automatically generated during digitization. Let the population of each individual region is as given in table 3.

Table 3: Population of the regions shown in figure 2

Region ID	Population
0	300
1	100
2	310
3	250
4	275
5	270
6	260
7	395
8	380
9	140
10	390
11	150
12	356
13	210
14	330
15	180
16	200
17	180
18	275
19	230
20	290
21	210
22	396

Thus the total population of the entire area considered is 6077. Let 3 segments has to form. The population of each segment will be approximately $6077/3 \approx 2025$, in order to make a uniform distribution of population among the segments. As it is not possible to touch the exact figure 2025 (because constituent regions could not be sub divided), so an amount of deviation factor has been associated, varying with application area. A deviation factor equal to average population (rounded upto nearest multiple of 100), which is ≈ 300 for the present example, has been set. Thus the population of each segment should lie within the range 2025 ± 300 = 1725 to 2325.

The following three segments are resulted, as shown in table 4, after application of the first phase of the method.

Table 4: Population of the segments formed formed after first phase of Abaft Proliferation Method

Segments	Population
0-1-2-3-11-12-13-14-17-18-22	2857
4-5-6-7-8-9-10-15-16	2490
19-20-21	730

The above segments (Table 4) are formed by considering only inter-region distances, without considering the population of the segments. It is clear that, among these three segments none is holding the required population (1725 to 2325). Here the second phase of the method comes onto the crease, which takes into account the associated attribute data (population). On completion of these second phase, the final segments achieved are shown in table 5.

Table 5: Population of the segments finally formed after by Abaft Proliferation Method

Segments	Population
0-1-2-3-11-12-13-14-17	2186
4-5-6-7-8-9-10	2110
15-16-18-19-20-21-22	1781

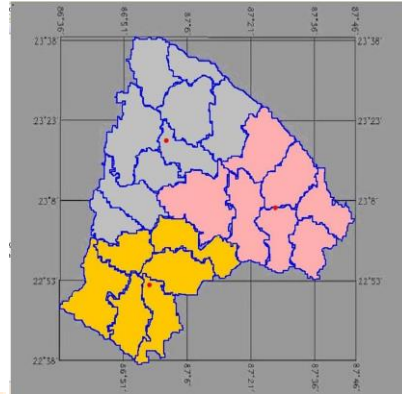The appearance of the segments are depicted in figure 3.

Fig. 3: Appearance of the segments formed by Abaft Proliferation Method

IV. CONCLUSIONS

For any country, especially for the rapidly developing third world countries like India, to meet the current need of its developing civilization, a numerous number of new constructions are needed. However the million dollar question is "Where"? To find the suitable most location for these new constructions is not easy enough. For any democratic country like India, election is supposed to be the backbone of the democracy. If Indian scenario is considered then it could be found that, taken into consideration different elections (like General or Loksabha, Assembly or Bidhansabha, Municipality, Panchayat etc.), election occurs almost once (sometimes even more) a year. Due to its high population, it is very much cumbersome and tedious to assign voters ward-wise into different polling station. The work becomes more difficult due to the fact that, the voters should be assigned into different polling stations in a manner so that they have not to travel much. Otherwise, people will not be interested to cast their votes, which in turn will weaken the pillar of democracy. In spite of setting up of temporary utility stations like this, there are a lot of situations which demands same while constructing permanent utility service centers like setting up of new banks, health centers, electricity offices etc. The present technique is a solution of this problem, which automatically allocates customers into respective service stations in a uniform way, making the job fast and simple. Moreover, a graphical representation increases the readability of the output generated.

REFERENCES

[1] Mohammad Aslani and A. Alesheikh Ali. Site selection for small gas stations using GIS. Scientific Research and Essays, 6(15):1361–3171, August 2011.

[2] S. Avdimiotis and E. Christou. GIS applications in tourism planning: A tool for sustainable development involving local communities. Journal of Environmental Protection & Ecology, 5(2):457–468, 2004.

[3] Heba Ayeldeen, Aboul Ella Hassanien, and Aly Aly Fahmy. Fuzzy clustering and categorization of text documents. In Proceeding of 13th International Conference on Hybrid Intelligent Systems (HIS 2013), pages 262–266. IEEE, October 2013.

[4] Tuul B. GIS BASED LAND INFORMATION SYSTEM FOR MANDAL SOUM, SELENGE AIMAG OF MONGOLIA . Technical report, GTZ, Land Management and Fiscal Cadastre project, Government building 12, ALAGCaC, Ulaanbaatar, Mongolia, June 2015. Accessed on 24 June, 2015.

[5] T. M. Bahaire and E. White. The application of geographical information systems (GIS) in sustainable tourism planning: a review. International Journal of Sustainable Tourism, 7(2):20–27, 1999.

[6] E. Balagurusamy. Numerical Methods. Tata McGraw-Hill Publishing Company Limited, New Delhi, second reprint edition, 2000.

[7] Maria-Florina Balcan, Yingyu Liang, and Pramod Gupta. Robust Hierarchical Clustering. Journal of Machine Learning Research, 15:4011–4051, 2014.

[8] Toju Francis Balogun, Henry Okeke, and Christain Ifanyin Chukwukere. Crime Mapping in Nigeria Using GIS. Journal of Geographic Information System, 6:453– 466, October 2014.

[9] Arpit Bansal, Mayur Sharma, and Shalini Goel. Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining. International Journal of Computer Applications, 157(6):35–40, January 2017.

[10] J. Beedasy and D. Whyatt. Diverting the tourists: A Spatial Decision-Support System for Tourism Planning on a Developing Island. International Journal of Applied Earth Observations and Geoinformation, Elsevier, 1(3):163–174, 1999.

[11] Steven Bell. An Overview of Optimal Graph Search Algorithms for Robot Path Planning in Dynamic or Uncertain Environments. Technical report, Oklahoma Christian University, March 2010.

[12] J. L. Bentley. Multidimensional binary search trees used for associative searching . Communications of the ACM, 18(9):509–517, 1975.

[13] H. Berger, M. Dittenbach, D. Merkl, A. Bogdanovych, S. Simoff, and C. Sierra. Opening new dimensions for eTourism. Virtual Reality, 11:75–87, 2007.

[14] S. Bertazzon, G. Crouch, D. Draper, and N. Waters. GIS Applications in Tourism Marketing: Current Uses, an Experimental Application and Future Prospects. Journal of Travel & Tourism Marketing, 6(34):35–59, 1997.

[15] J. C. Bezdek and Full W. Ehrlich R. FCM: The Fuzzy c-Means Clustering Algorithm. Computers and Geosciences, 10(2-3):191–203, May 1984.