Spatial Clustering with Obstacles Constraints Based on Genetic Algorithms and K-Medoids

Xueping Zhang [†,††,†††], Jiayao Wang [†], and Fang Wu[†]

[†] Institute of Surveying and Mapping, PLA Information Engineering University, Zhengzhou 450052, China.
[††] School of Computer Science and Engineering, Henan University of Technology, Zhengzhou 450052, China.
[†††] Geomatics and Applications Laboratory, Liaoning Technical University, Fuxin 123000, China.

Summary
Spatial clustering has been an active research area in Spatial Data Mining (SDM). Many methods on spatial clustering have been proposed in the literature, but few of them have taken into account constraints that may be present in the data or constraints on the clustering. In this paper, we discuss the problem of spatial clustering with obstacles constraints and propose a novel spatial clustering method based on Genetic Algorithms (GAs) and K-Medoids, called GKSCOC, which aims to cluster spatial data with obstacles constraints. The GKSCOC algorithm can not only give attention to higher local constringency speed and stronger global optimum search, but also get down to the obstacles constraints and practicalities of spatial clustering. The results on real datasets show that the GKSCOC algorithm performs better than the IKSCOC algorithm in terms of quantization error.

Key words:
Spatial Clustering, Genetic Algorithms, K-Medoids Algorithm, Obstacles Constraints.

1. Introduction

Spatial clustering has been an active research area in the data mining community. Spatial clustering is not only an important effective method but also a prelude of other task for Spatial Data Mining (SDM). As reported in surveys on data clustering, clustering methods can be classified into Partitioning approaches, Hierarchical methods, Density-based algorithms etc. As pointed out earlier, these techniques have focused on the performance in terms of effectiveness and efficiency for large databases. However, few of them have taken into account constraints that may be present in the data or constraints on the clustering. These constraints have significant influence on the results of the clustering process of large spatial data.

Spatial clustering with constraints has two kinds of forms [1]. One kind is spatial clustering with obstacles constraints, such as bridge, river, and highway etc. whose impact on the result should be considered in the clustering process of large spatial data. The other kind is spatial clustering with handling operational constraints [2], it consider some operation limiting conditions in the clustering process. In this paper, we only discuss spatial clustering with obstacles constraints.

Since K.H.Tung put forward a clustering question COE (Clustering with Obstacles Entities) in [3], a new studying direction in the field of clustering research have been opened up. To the best of our knowledge, only three clustering algorithms for clustering spatial data with obstacles constraints have been proposed very recently: COD-CLARANS [3], AUTOCLUST+ [4], and DBCluC [5]-[8]. Although these algorithms can deal with obstacles in clustering, many questions exist in them. COD-CLARANS algorithm only gives attention to local constringency. AUTOCLUST+ algorithm builds a Delaunay structure to cluster data points with obstacles costly and is unfit for a large number of data. DBCluC cannot run in large high dimensional data sets etc.

In order to overcome the disadvantage of Partitioning-based approach, which only gives attention to local constringency, and keep the advantage of fast constringency at the same time, we propose a novel Spatial Clustering with Obstacles Constraints based on Genetic algorithms (GAs) and K-Medoids, called GKSCOC. The GKSCOC algorithm cannot only give attention to higher local constringency speed and stronger global optimum search, but also get down to the obstacles constraints and practicalities of spatial clustering. The results show that the GKSCOC algorithm performs better than the IKSCOC algorithm in terms of quantization error.

The remainder of the paper is organized as follows. Section 2 introduces spatial clustering based on GAs. Spatial Clustering with Obstacles Constraints based on K-Medoids is discussed in Section 3. Section 4 presents Spatial Clustering with Obstacles Constraints based on GAs and K-Medoids. The performance of GKSCOC in comparison with the standard K-Medoids and GAs are showed in Section 5, and Section 6 concludes the paper.

Manuscript received October 15, 2006.
Manuscript revised October 28, 2006.

information for authors, please refer to [1].

2. Spatial clustering based on GAs

Genetic algorithms are an efficient parallel and near global optimum search method based on nature genetic and selection combining the concept of survival of the fittest with a structured interchange, but uncertain component of the information [9]. GAs imitate natural selection of the biological evolution and use the technology of searching population (set) to produce a new generation population and evolve into the optimum state progressively by exerting a series of genetic operators such as selection, crossover and mutation etc. GAs can not only automatically achieve and accumulate the knowledge about the search space during the search process, and adaptively control the search process to approach a global optimal solution, but also perform surprisingly well in highly constrained problems, where the number of "good" solutions is very small relative to the size of the search space. Furthermore, GAs can provide almost the better solution in a shorter time, including complex problems to solve by traditional methods.

Clustering purpose is to divide a given group of objects in a number of groups (clusters), in order that the objects in a particular cluster would be similar among the objects of the other ones. This technique tries to solve how to distribute n object in k clusters according to the minimization of some optimization criterion additive over the clusters. Once the optimization criterion is selected, the clustering problem is to provide an efficient algorithm in order to search the space of the all possible classifications and to find one on which the optimization function is minimized.

The basic algorithm of spatial clustering based on GAs can be described as follows. Divide an individual (chromosome) into n part and each part is corresponding to the classification of a datum element. The optimization criterion is defined by a Euclidean distance among the data frequently, and the initial population is produced at random. Its genetic operators are similar to standard GA's. This method can find the global optimum solution and not influenced by an outlier, but it only fits for the situation of small data sets and classification number [9]-[11].

3. Spatial Clustering with Obstacles Constraints Based on K-Medoids

Partitioning-base algorithm divides n objects into $k(k < n)$ parts, and each part represents one cluster. There are there typical types of partitioning-based algorithm: K-Means, K-Medoids and CLARANS. K-Means takes the average value of a cluster as the cluster centre. While adopting this algorithm, a cluster center possibly just falls on the obstacle (Fig.1), and it cannot be implemented in reality. On the other hand, K-Medoids takes the most central object of a cluster as the cluster centre, and the cluster center cannot fall on the obstacle. In view of this, K-Medoids algorithm is adopted in this paper.

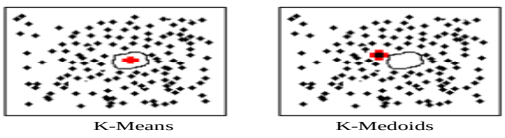

Fig.1. K-Means vs. K-Medoids

3.1 K-Medoids Algorithm

K-Medoids algorithm selects the most central object of a cluster as the cluster centre. The basic thought of K-Medoids algorithm is as follows. A representing object is selected for each cluster at random and remaining objects are distributed to the nearest cluster according to their distance from the representing object. In order to improve the quality of the cluster, the representing object is replaced with the other object repeatedly. The clustering quality is estimated by an object function. Square-error function is adopted here, and its definition is:

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} (d(p, c_i))^2 \quad (1)$$

Here, c_i is the cluster centre of cluster C_i, $d(p,q)$ is the direct Euclidean distance between the two points p and q.

The standard K-Medoids algorithm can be described as follows.

1. Select k objects to be cluster centers at random;
2. Distribute remain objects to the nearest cluster center;
3. Calculate current $E = E$;
4. Do {
5. Select a not centering point to replace the cluster center c_i randomly;
6. Distribute objects to the nearest cluster center;
7. If $E <$ current E then form new cluster centers;
8.} While (E changed).

3.2 Spatial Clustering with Obstacles Constraints Based on K-Medoids

In order to handle obstacles constraints, the obstructed distance function $d'(p,q)$ is defined by the shortest distance between the two points p and q which cannot be

cut off by any obstacle. Accordingly, criterion function is revised:

$$E' = \sum_{i=1}^{k} \sum_{p \in c_i} (d'(p, c_i))^2 \quad (2)$$

Computing obstructed distance can be divided 3 steps as follows [3]:

1. If p and q are visible to each other then $d'(p,q) = d(p,q)$;
2. Supposing the visible obstacles vertices of p is $vis(p) = \{v_1, v_2, ..., v_n\}$ (Fig.2), $d(p, v_i)$ is the direct Euclidean distance between the two points p and v_i, $1 \le i \le n$;
3. Choosing v_m in $vis(p)$ to minimize $d(p, v_m) + d'(q, v_m)$ and then $d'(p,q) = d(p, v_m) + d'(q, v_m)$.

Further explanations and details on how to calculate obstructed distance $d'(p,q)$ can be found in [3].

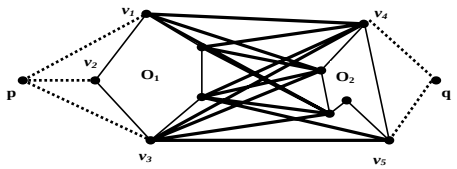

Fig.2. Visibility graph

3.3 Improved Spatial Clustering with Obstacles Constraints Based on K-Medoids

In order to improve the efficiency of whole algorithm, the method of Improved KSCOC (IKSCOC) is adopted as follows.

1. Select k objects to be cluster centers at random;
2. Distribute remain objects to the nearest cluster center;
3. Calculate E';
4. Do {let current $E = E'$;
5. Select a not centering point to replace the cluster center c_i randomly;
6. Distribute objects to the nearest center;
7. Calculate E; if $E >$ current E then go to 5 otherwise continue;
8. Calculate E';
9. If $E' <$ current E then form new cluster centers;
10.} While (E' changed).

While IKSCOC still inherits two shortcomings. One is that selecting initial value randomly may cause different results of the spatial clustering and even have no solution. The other is it only gives attention to local constringency and is sensitive to an outlier.

4. Spatial Clustering with Obstacles Constraints Based on GAs and K-Medoids

In order to overcome the disadvantage of partitioning approach which only gives attention to local constringency, and keep the advantage of GAs which has stronger global optimum search at the same time [10], we propose a novel Spatial Clustering with Obstacles Constraints based on GAs and K-Medoids (GKSCOC) as follows:

1. Initialization: initialize genetic parameter, produce an initial population $P(0)$;
2. Calculate fitness of every individual in the population $P(t-1)$;
3. Select $P(t-1)$ to get a new population $P(t)$;
4. Cross $P(t)$;
5. Mutate $P(t)$;
6. Optimize new individuals by K-Medoids;
7. If the iterative times get to greatest generation, or the value of J between $P(t-1)$ and $P(t)$ is in threshold, continue otherwise go to 2;
8. Output.

4.1 Objective function

Where, objective function is defined as follows:

$$f(S_i) = \frac{1}{J_i} \quad (3)$$

$$J_i = \sum_{k=1}^{m} \sum_{x_j \in C_k} d'(x_j, P_k) \quad (4)$$

Thus the lower J_i is, the higher the fitness value is.

4.2 Encoding schemes

An individual (chromosome) is encoded as the following string:

$S = P_1 P_2 \cdots P_m$

Where, $P_j (j = 1, \cdots, m)$ are the cluster centers.

And the nearest neighbor rule is defined as follows:

$$d'(x_i, P_j) = \min_{k=1,\cdots,m} (d'(x_i, P_k)) \quad (5)$$

If the distance between $x_i (i = 1, 2, \cdots, n)$ and $P_j (j = 1, \cdots, m)$ satisfy (5), then x_i is contained in the cluster j.

4.3 Genetic operators

During *selection* and reproduction, pairs of individuals are chosen from the population according to their fitness. Here, we consider roulette wheel selection (also called fitness proportional model).

The *crossover* operator is used to combine the pairs of selected individuals to create new individuals that potentially have a higher fitness than either of their parents. We design the following crossover operator based on nearest neighbor gene match.

Supposing $S_1 = P_1^{(1)} P_2^{(1)} \cdots P_m^{(1)}$ and $S_2 = P_1^{(2)} P_2^{(2)} \cdots P_m^{(2)}$ ar e individuals to cross, for each $P_i^{(1)}$ in S_1, select $P_j^{(2)}$ in S_2 such that $P_j^{(2)}$ is the closest one to $P_i^{(1)}$, and mate $P_i^{(1)}$ and $P_j^{(2)}$. During gene mating, the gene that has already mated would not participate in the follow mating any longer. In this way, the genes of S_1 and S_2 are mated finally. And then reconfigure S_2 to get S_2^* according to the order in which the genes mated. For S_1 and S_2^*, cross by selecting crossover points at random to get new individuals S_1' and S_2'.

The *mutation* operator is applied to every individual resulting from the crossover process. When mutation is applied, each character of the individual has a low probability (e.g., typically 4/1000) of being changed to another random value of the same type and range. Here, mutate according to gene location. Every cluster centre on gene location mutates at random by MR, and the mutant gene is replaced by the one who is chosen randomly.

4.4 Individuals Optimization

For new individuals produced in every evolving, optimize them according to the following one:

1. Cluster individuals according to (5);
2. Calculate new cluster centers by IKSCOC.

4.5 Individuals Adjustment

In K-Medoids optimizing and individuals crossing, normal individuals might appear. In order to avoid this situation, adjust normal individuals according to the following: if a certain cluster C_i is empty then select an object from the non-empty cluster which is closest to the cluster C_i, and put it into C_i. Repeat this course until there is no empty cluster again.

5. Results and Discussion

We have made experiments on the cluster of residential spatial data points with river obstacles 8 times separately by K-Medoids, IKSCOC and GKSCOC. The number of clusters k= 4, M =50, $P_c = 0.6$, $P_m = 0.001$, the greatest generation is 100, and ε =0.001 (kilometer). Fig.3, Fig.4 and Fig.5 are respectively the results of K-Medoids, IKSCOC and GKSCOC. Obviously, the results of the clustering illustrated in Fig.4 and Fig.5 both have better practicalities than that in Fig.3, and the one in Fig.5 is superior to the one in Fig.4.

Fig.6 and Fig.7 displayed respectively the map of K-Medoids and GKSCOC showing the 4 clusters of city spatial data around the Yellow River. Obviously, the map in Fig.7 displayed better practicalities than that in Fig.6.

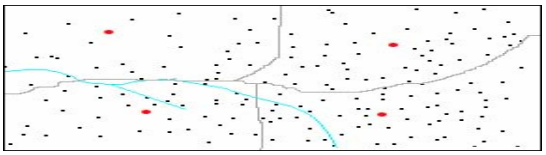

Fig.3. Result of K-Medoids

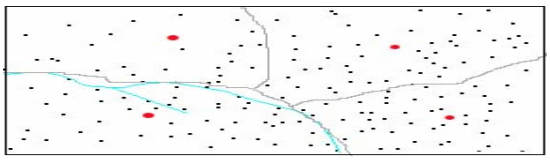

Fig.4. Result of IKSCOC

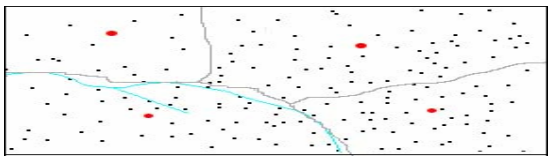

Fig.5. Result of GKSCOC

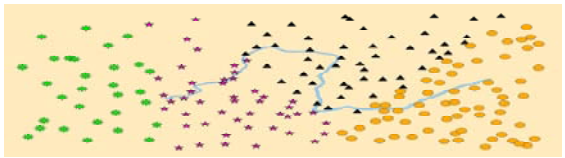

Fig.6. MAP of K-Medoids showing the 4 clusters of city spatial data around the Yellow River

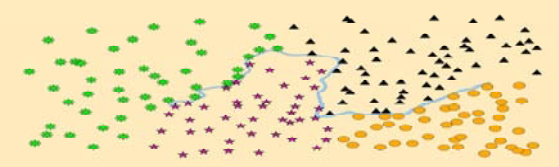

Fig.7. MAP of GKSCOC showing the 4 clusters of city spatial data around the Yellow River

Table 1 is the value of J showed in every experiment. It is showed that IKSCOC is sensitive to initial value and it constringes in different extremely local optimum points by starting at different initial value while GKSCOC constringes nearly in the same optimum points each time. Therefore, we can draw the conclusion that GKSCOC has stronger global constringent ability comparing with KSCOC; and GKSCOC has not only considered high local constringent speed but also kept good global constringent characteristic, but the drawback of this method is a comparatively slower speed in spatial clustering.

Table 1. Value of J showed in experiments (unit kilometer)

Algorithms	***First***	***Second***	***Third***	***Forth***
IKSCOC	3665.1	3881.5	3720.5	4081.3
GKSCOC	3691.6	3691.3	3690.4	3990.0
	Fifth	***Sixth***	***Seventh***	***Eighth***
IKSCOC	3981.4	3539.4	4021.6	3789.5
GKSCOC	3690.8	3690.5	3690.6	3690.6

6. Conclusions

In this paper, we discussed the problem of spatial clustering with obstacles constraints and propose a novel GKSCOC based on GA and K-Medoids. The comparison proves that our method can not only give attention to higher local constringency speed and stronger global optimum search, but also get down to the obstacles constraints and practicalities of spatial clustering. The results of the experiments on real datasets show that the GKSCOC algorithm performs better than the IKSCOC algorithm in terms of quantization error. The drawback of GKSCOC algorithm is a comparatively slower speed in clustering. But its achievements will have more practical value and extensive application prospect.

Acknowledgments

This work is partially supported by the Natural Sciences Fund Council of China (Number: 40471115）, the Natural Sciences Fund of Henan（Number:0511011000）and the Open Research Fund Program of the Geomatics and Applications Laboratory, Liaoning Technical University (Number: 2004010).

References

[1] A.K.H.Tung, J.Han, L.V.S.Lakshmanan, and R.T.Ng. “Constraint-Based Clustering in Large Databases,” In Proceedings of the International Conference on Database Theory (ICDT'01) [C], London, U.K., 2001. pp. 405-419.

[2] A.K.H.Tung, R.T.Ng, L.V.S.Lakshmanan, and J.Han. “Geo-spatial Clustering with User-Specified Constraints”. In Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000) [C], in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000.

[3] A.K.H.Tung, J.Hou, and J.Han. “Spatial Clustering in the Presence of Obstacles”, In Proceedings of International Conference on Data Engineering (ICDE'01) [C], Heidelberg, Germany, April, 2001. pp. 359-367.

[4] V.Estivill-Castro and I.J.Lee. “AUTOCLUST+: Automatic Clustering of Point-Data Sets in the Presence of Obstacles”. In Proceedings of the International Workshop on Temporal, Spatial and Spatial-Temporal Data Mining [C], Lyon, France, 2000. pp. 133-146.

[5] O.R.Zaïane and C. H.Lee. “Clustering Spatial Data When Facing Physical Constraints”. In Proceedings of the IEEE International Conference on Data Mining (ICDM'02) [C], Maebashi City, Japan, 2002. pp. 737-740.

[6] X.Wang and H.J.Hamilton. “DBRS: A Density-Based Spatial Clustering Method with Random Sampling.” In Proceedings of the 7th PAKDD [C], Seoul, Korea, 2003. pp. 563- 575.

[7] X.Wang, C.Rostoker and H.J.Hamilton. “DBRS+: Density-Based Spatial Clustering in the Presence of Obstacles and Facilitators.” ftp.cs.uregina.ca/Research/Techreports/2004-09.pdf, 2004.

[8] X.Wang and H.J.Hamilton. “Gen and SynGeoDataGen Data Generators for Obstacle Facilitator Constrained Clustering.” ftp.cs.uregina.ca/Research/Techreports/2004-08.pdf, 2004.
[9] V.Fernández, R.García Martínez, R.González, and L.Rodriguez. “Genetic Algorithms Applied to Clustering.” In Proceedings of the 1997 Winter Simulation Conference [C], 1997. pp. 1307-1314.
[10] D. Doval, S.Mancoridis, and B.S.Mitchell. “Automatic Clustering of Software Systems using a Genetic Algorithm.” In Proceedings of the International Conference on Software Technology and Engineering Practice (STEP) [C], Pittsburgh, PA, 1999. pp. 73-91.
[11] V.Estivill-Castro and A.T. Murray. “Spatial Clustering for Data Mining with Genetic Algorithms”. Technical Report FIT-TR-97-10, 1997. http://sky.fit.qut.edu.au/TR/techreports
[12] J.Wang, X.Zhang, and H.Zhou. “A Genetic K-Means Algorithm for Spatial Clustering.” Computer Engineering [J] (in Chinese). No.3, 2006. pp. 188-190.

Xueping Zhang received the B.S. and M.S. degrees in Computer Science from ZhengZhou University in 1990 and 2001, respectively. After working as an assistant professor (from 1996) in the School of Computer Science and Engineering, Henan University of Technology, she has been an associate professor at Henan University of Technology since 2002. Her research interest includes Data Mining, Artificial Intelligence and Software Engineering, and their application to GIS. Now, she is a PhD student in GIS in the Institute of Surveying and Mapping, PLA Information Engineering University.

Jiayao Wang is a member of Chinese Academy of Engineering, and he is a professor at Institute of Surveying and Mapping, PLA Information Engineering University. His research interest includes theoretical cartography, spatial information system, and so on.

Fang Wu is a professor at Institute of Surveying and Mapping, PLA Information Engineering University. Her research interest includes theoretical cartography, GIS, and so on.