THE SPATIAL ANALYSIS OF CLUSTERING BASED ON GENETIC ALGORITHMS

Hongyan DENG, Fang WU, Chang YIN

Institute of Surveying and Mapping, Information Engineering University, Zhengzhou, China 450052
zb3535@sina.com

Commission II, IC WGⅡ/Ⅳ

KEY WORDS: Genetic algorithms, Spatial analysis, Clustering

ABSTRACT:

On the basis of analyzing the existing methods of the spatial analysis of clustering, considering the basic principles and characters of genetic algorithms, this paper provides a new method of the spatial analysis of clustering based on genetic algorithms and discusses the keys. The experimental output indicates that this method can keep general character of distribution and get good output.

1. INTRODUCTION

Spatial analysis is a very familiar aspect to the peoples who study in the field of dialect geography. Since the appearance of map, Peoples always carry through all kinds of spatial analysis by self-conscious or unconscious. For example, the measure of the distance, the azimuth between two features in the map, and doing tactical research and strategic decision-making by using map and so on. Because spatial analysis can distill and transmit spatial information, it has become primary functional character of geographical information system which is knows from ecumenical information system.

Clustering is a very important aspect in spatial analysis. It describes spatial variables and the characters of spatial objects from the general, total viewpoint, and it reflects the information of orientation of cluster. Its aim is to analyze the multitudinous character of spatial objects and divide one cluster into several different clusters, to discover some geographical characteristic or do the base of other analysis. Spatial analysis of clustering can be thought as a kind of general optimization algorithms substantially.

Genetic algorithms (GA) is on the basis of natural selection and genetic theory. It is such a search algorithms that connects the rules of survival of the fittest with the stochastic commutative information of chromosomes. It can do general parallel search and it is simple, rapid and robust. On the basis of analyzing the exiting algorithms of the spatial analysis of clustering, considering the basic principles and characters of genetic algorithms, this paper provides a new method of the spatial analysis of clustering based on genetic algorithms.

2. INTRODUCTION OF THE METHODS OF SPATIAL CLUSTERING

The spatial analysis of clustering is one kind of clustering which is based on the geometrical position of spatial data. So, the spatial clustering character of points is based on this foundation mainly. Generally speaking, the spatial distance between points is the most primary statistical value that decides the clustering character of points. Only the points that the distances among them are close enough can be accredited one cluster, other wise, they will be accredited different clusters. So, in spatial clustering, the rule in common use is distance. In this article, we will use this rule too.

Spatial clustering can adopt different methods. These methods can be boiled down to three genus: systemic clustering, stepwise decomposition and distinguish—clustering.

2.1 Systemic Clustering

First of all, the n points are assumed as n clusters, then, we converge them stepwise. In such process of clustering, the number of clusters will become less and less till it reaches an appropriate value. This process is named systemic clustering. The idiographic process is that first, the two clusters, which the distance between them is the nearest, are converged one cluster, we will get n-1 clusters, then, the distances between these n-1 clusters should be recalculated and we will select the two clusters which distance is the nearest to converge and get n-2 clusters ……till the number of clusters reaches a perfect value.

Because only the statistical value of those clusters that process coalition need be recalculated, the distances between the other clusters needn't be recalculated. So, the quantum of calculation is small and the process is simple. This kind of method has been comparative mature and applied widely.

2.2 Stepwise Decomposition

In the process of clustering, above of all, the n points are assumed as one cluster, then, we will decompose them stepwise. So, the number of clusters will be more and more by the process till it researches an appropriate value. This process of clustering is named stepwise decomposition.

Many aspects need be considered in the stepwise decomposition and this kind of method is not mature enough. So, it is not used widely in practice.

2.3 Distinguish—Clustering

Umpty centers of clustering should be appointed firstly. Then, the distances between every point and these centers should be compared to decide the adscription of every pint. This kind of method is named distinguish-clustering. The key to this method is how to appoint centers. Peoples are studying and mending these methods ceaselessly. So distinguish-clustering is accepted and used widely by people step by step.

Systemic clustering and stepwise decomposition are all based on the partial consideration. In these processes, preserving the general characteristic of distributing does not be considered. In the process of distinguish—clustering, if the centers of clustering accord with the general characters of distributing, The output will preserve the general characters of distributing well. But this method depends on the way of the selection of centers greatly. So, We must find a new method that considers the general characters of distributing. Finding centers of clustering which has this character will be a good shortcut.

3. THE BASE PRINCIPLES AND CHARACTERS OF GA

Genetic algorithms was brought forward by professor T.Holland of American Michigan University when he and his students studied natural adaptive systems in the latter 1950s and early 1960s. Now, it has frequently run to a comprehensive applied, effective general optimization method. The GA combined the principles of biologic evolution 、 the technology of optimization and the technology of computer and it exploits a new applied field.

GA has three elementary operators: selection operator 、 crossover operator、 mutation operator.

GA、 simulated annealing and NN go by the name of three unclassical numeric optimization algorithms. These algorithms conquer some disadvantage of traditional algorithms, and have obvious advantages that aim at some complicated systems. Just about because the GA make full use of the ideas of biologic evolution and heredity, It has many characters that are different from the traditional optimization algorithms：

1. GA uses the coding of decision-making values as operation object. It appears its particular superiority in allusion to some optimization problems that do not have or is difficult to have numeric concepts;
2. GA uses aim function as contractive information directly;
3. GA uses research information of multi-points.
4. GA uses the research technology of probability.

GA provides one all-purpose framework that is used to solve complicated systems. It does not depend on the idiographic field of problems and has strong the character of transfer。So it is used in many subjects widely. This article will combine the GA with distinguish-clustering, and provide one kind of method that is the spatial analysis of clustering based on genetic algorithms.

4. THE SPATIAL ANALYSIS OF CLUSTERING BASED ON GA

Combining the ilka characters of GA and distinguish—clustering, It is not difficult to find that the key to distinguish—clustering is the selection way of the centers of clustering, and GA has the characters of general search. So, We will use GA to find the centers of clustering that holds general characters automatically. Then, Use distinguish-clustering to judge the adscription of other points. This process must get output that accords with general characters of distributing.

Several main problems should be solved in this case:

4.1 Coding

Coding is the first problem to be solved as using GA and it is a key step as designing GA too. Besides the method of coding decides the tactic form of every chromosome (the idiographic value of one chromosome), it decides the method of decoding that the individuals transform from the gene of search space to the representation of explain space. And it will infect the method of selection operator、 crossover operator、 mutation operator.

When we search the centers of clustering from a cluster of points that need to do the spatial analysis of clustering, every point will be a center or not. So Using the method of coding to represent every selection output will meet such a problem: if the structure of coding is too complicated, the calculation will be very complicated when the number of points is too large. So, based on the principle of simple and easy, in the process of coding, we select the binary code that can be coded, decoded, crossed easily.

The idiographic coding method is introduced as follows. The length of coding is the number of points, the value of allele reflects whether the point in the same situation is selected as the center of clustering. If the value is 1, the point is selected; if the value is 0, it isn't selected. For example, if the aggregate of points is P{p1,p2,p3……p8},if the chromosome is 10011001, this show that p1,p4,p5 and p8 are the centers of clustering.

4.2 Fitness Function

In GA, the genetic probability of individual is made certain by the fitness of individual. In general speaking, if its fitness is big, the genetic probability will be big too. How to make certain fitness function has great effect on the capability of GA.

Fitness function is contacted with the requests of question nearly. In this case, we design the fitness function as follows:

$$F(t)=SUM(P) \tag{1}$$

where SUM(P) is the number of the selected centers of clustering. Considering the accuracy, we hope SUM(P) will be the least .

4.3 The Operators of GA

The operators of GA is mainly crossover operator and mutation operator. Now, we adopt the usual single crossover operator (Figure 1). But we do not adopt the usual mutation operator, we use partial optimum selection instead. This action will guarantee that every chromosome to be dealt with will be one reasonable solution. In this case, we must guarantee that every point that is not center must vest in one of the points that are centers of clustering and the distance between two centers will not be too close. The idiographic steps is introduced as follows.

```
A=1 1 1 0 0 0 0 1 1 1 0|1 0 1 0 1
B=0 0 0 0 0 0 0 0 1 0 1|0 1 1 1 1
```

（a）Two parents

```
A1=1 1 1 0 0 0 0 1 1 1 0 0 1 1 1 1
B1=0 0 0 0 0 0 0 0 1 0 1 1 0 1 0 1
```

（b）Two sun

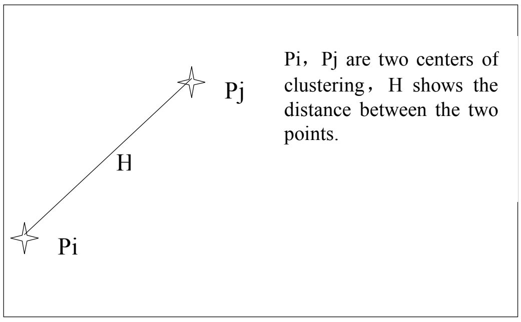
Figure 1. Single crossover operator

4.3.1 The Decoding Disposal of Chromosome

The chromosome that will be done by partial optimum selection is decoded firstly. The centers of clustering will be selected from the original points.

4.3.2 The Selection of Clustering Center

According to the fact position of every clustering center, Judge if it should be one clustering center (Figure 2.). Supposing pi and pj are two clustering centers. If the distance between them is shorter than the limit, it shows that pi and pj is too near. The value of the pi allele will change from 1 to 0 and pi will be not clustering center.

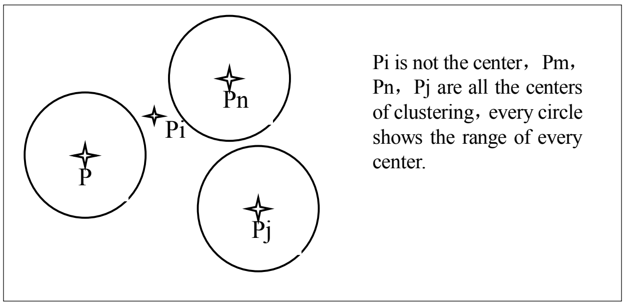


Pi，Pj are two centers of clustering，H shows the distance between the two points.

Figure 2. Judgement of reasonable degree of the center of clustering

4.3.3 The Judge of the Other Points

Judge if every point that is not center has its ascription. If pi doesn't belong to every clustering center according to the rule of distance, it will become clustering center and the allele value of chromosome will change 1(Figure 3.).



Pi is not the center，Pm，Pn，Pj are all the centers of clustering，every circle shows the range of every center.

Figure 3. Judge of the point that is not center

4.3.4 Replacing Chromosome

The old chromosome will be replaced by the repaired chromosome.

Connecting GA with partial optimum selection as above to use heuristic information with the correlative knowledge is one of the main strategies to promote the efficiency of GA. It can void wasting time on the solution that is not reasonable. It will promote the efficiency of system greatly and impenetrate the whole GA.

4.4 The Parameters of GA

The parameters of GA mainly includes population size M, crossover rate Pc and probability of mutation Pm. These parameters have great effect on GA and need to be selected carefully. But these parameters are got by the way of experiment mainly.

By the compare of experiment, in this article, population size is 50, Pc is 0.8 and Pm is 0.

4.5 The Selection Operator

During the process of biologic heredity and natural evolutionary, the species that can adapt to the evolution well will have more chances to inherit, otherwise, they will have less chances. Imitating this process, the GA will use selection operator to do the operation that only those organisms best adapted to existing conditions are able to survive and reproduce.

The selection operators of GA are various. The operators which be used usually are roulette wheel selection and tournament selection and so on. In this article, we select tournament selection. This selection method will make the individual which has great fitness has bigger chance to survive and it only use the opposite fitness value which doesn't pro rate with the size of value as selection rule and it will avoid the effect of super individual.

4.6 Washed-up Conditions

This article will adopt two washed-up conditions: one condition is the number of genetic generation. It shows that the GA will stop if it reaches the number of genetic generation. The other condition is that average fitness is compared with the biggest fitness, if the discrepancy between average fitness and the biggest fitness is in the limit range, the GA will stop. It shows that the GA will not get better output if the process goes on. Combine the two conditions, the GA will stop no matter what it meets which condition.

According to the steps hereinbefore, take these pivotal keys to solve GA into the framework of GA. We will get the centers of clustering, then, according to the rule of distance, judge the ascription of every point that is not center cluster.

5. CONCLUSION

This article provides the method of the spatial analysis based on GA, analyze its feasibility in theory and validate it technically. The experiment indicates that the spatial analysis of clustering based on genetic algorithms preserves the general characters of distributing well(Figure 4.).

The spatial analysis of clustering is only one small aspect of the applications of GA in spatial analysis. Many problem of spatial analysis will be thought as or changed to the optimum problem. So, we need to do further work.

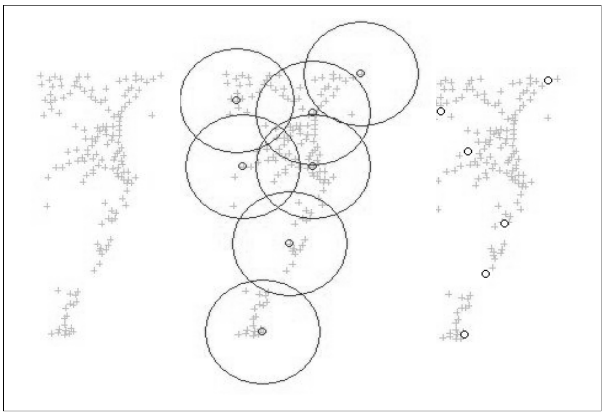


Figure 4. The output of the spatial analysis of clustering based on GA

REFERENCES

Bader, M. and Weibel, R., 1997, Detection and resolving size and proximity conflicts in the generalization of polygonal maps, *Proceedings of the 18th International Cartographic Conference*, pp.1525-1532.

GUO Renzhong, 1997. *Spatial Analysis.* WuHang, China, pp.86-103.

Kim Lowell, Gold C, 1995. Using a Fuzzy surface-based cartographic Representation to Decrease Digitizing Efforts for Natural Phenomena. *Cartography and Geographic Information Systems*, Vol.22, No.3, 1995

ZHOU Ming, SUN ZeDong, 1999. *Genetic Algorithms Theory and Applications*, Beijing , China.

ZHANG Wenliane, LIANG Yi, 2000. *The Mathematic Base of Genetic Algorithms,* Xi'an, China.

Weibel, R., 1995, Map generalization in the context of digital system. *Cartography and Geographic Information System*, Vol.22, No.4, pp.259-263.