A genetic algorithm for spatiotemporal cluster detection and analysis

Dajun Dai[1] and Tonny J. Oyana[2]

Department of Geography,
Southern Illinois University
1000 Faner Drive, MC 4514,
Carbondale, IL 62901-4514
Tel: +1 618-453-3022
Fax: +1 618-453-6465

1. Email : dljt@siu.edu
2. Email: tjoyana@siu.edu

Abstract

Although increased exploration of large-scale databases has provided the impetus for better detection and analysis of spatial clusters, there is slow progress in developing clustering algorithms for classifying space-time multidimensional attributes and space-time-attribute interactions. The objective of this study is to enhance the genetic algorithm for detecting clusters in spatiotemporal or more complex hyperspace. Our motivation is guided by the idea of representing the gene structure using an elliptic cylinder allowing the detection and analysis of spatial clusters that have space-time components or space-time-attribute dimensions. A more sophisticated gene representation through elliptic cylinders can model variable of spaces of two dimensions and greater. To assess the method, we employed a published real-world dataset with known spatiotemporal clusters of brain cancer incidence in New Mexico. Experimental results are compared with the results obtained from the very popular cluster detection method, Kulldorff's space-time scan statistic. The results indicate that the proposed method provides a better representation of clusters in space-time and space-time-attribute interaction at faster and more computationally efficient runtime than Kulldorff's method.

1. Introduction

Identifying clusters where there is higher-than-expected number of cases has been the subject of much research in geographic information science, epidemiology, criminology, and other fields (e.g. Snow 1855; Mantel 1967). In epidemiology and criminology, objects may be individual cases, aggregated into different administrative regions, or grouped into different time periods. Because the number of cases follows Poisson distribution, under the null hypothesis, the expected number of cases in each area is proportional to the size of its background population at risk if there are no covariates (Kulldorff 2005). Scientists therefore are interested in finding the reasons that can explain where spatiotemporal factors where higher-than-expected number of events occurs.

Although many methods have been developed for spatial clustering, there is still a great need to develop clustering algorithms accounting for space-time-attribute interactions with generalized assumptions. Space-time-attribute interactions may have hidden patterns (Turton et al. 2000). Pure spatial clustering methods tend to ignore totally these interactions. In addition, most algorithms assume clusters to be in circular shape, such as Geogrpahic Analysis Machine (GAM; Openshaw et al. 1987), Map Explore (MAPEX; Openshaw and Perrée 1996). This assumption is generalized in PROCLUDE (Conley et al. 2005) that clusters are represented by horizontal and vertical ellipses. However, geographic clusters may be elongated with some kind of orientations, which is previously unknown. Although most methods are good at detecting irregularly shaped clusters, such as DBSCAN (Ester et al. 1996), CHAMELEON (Karypis 1999), AUTOCLUST (Estivill-Castro and Lee 2000), they are limited in exploring the space-time-attribute interactions. Therefore, clustering methods that can detect diagonally elongated clusters in both space and time are still limited and have not been fully developed.

The objective of this study is to develop a clustering method accounting for space-time-attribute interactions using genetic algorithms (GA). GAs have been proven effective in searching spatial clusters, such as MAPEX, PROCLUDE, and Hobbs and Goodchild's method (Hobbs and Goodchild 1996). However, few efforts have been made to use GAs for spatiotemporal clustering. Our motivation is therefore guided by the idea of using elliptic cylinder search window to identify clusters in space-time-attribute interactions.

2. Proposed genetic algorithm for spatiotemporal cluster detection and analysis

2.1 Representation of individuals

In our proposed genetic algorithm, each individual is an elliptic cylinder search window with an elliptic base and with height corresponding to time. The elliptic base covers less than or equal to half the total study area. The height reflects any possible time interval of less than or equal to half the total study period.

Each individual has 7 parameters: centroid (x, y), semi-major axis (a), semi-minor axis (b), an orientation angle (θ) from the horizontal line to the major axis, starting time (Ts), and time interval (Tin). The time period ranges from Ts to Ts+Tin. A criticism may exist that parameters (a, b and θ) are redundant if the cluster is circular or close to circular. Given that the shapes of clusters are usually previously unknown in real world; the proposed method generalizes the assumption, thus providing a higher degree of freedom. One concern, however, is that the proposed algorithm is not very robust in detecting irregularly shaped clusters; this is mainly because our focus, at least for now, is to develop an algorithm that captures spatiotemporal factors.

2.2 Population initialisation

A population of n individuals is randomly generated. The encoding process defines low and high bounds to each parameter. Therefore, each individual elliptic cylinder search window is located within the study area and within the study period.

2.3 Fitness evaluation

Fitness evaluation is the key step to ensure that better fit individuals can survive into next generation. In health or crime studies, cases are assumed to be Poisson distributed with constant risk over space and time under the null hypothesis. In order to account for population at risk and relevant covariates, the fitness function is assigned as follows:

$$FitV = c - C\frac{p}{P}, \tag{1}$$

where c is the observed number of cases contained in an elliptic cylinder, p is the background population contained within the elliptic cylinder, C is the total number of cases observed in the study area and period, and P is the total population in the study area and period.

2.4 Selection

The selection operation selects individuals from the current population for genetic reproduction, crossover, and mutation operations. In the selection process, the fitness value of each individual is calculated. High-quality individuals have a higher chance to be copied into next generation. The selection operator is implemented using a stochastic universal sampling method (Goldberg 1989).

2.5 Crossover

The crossover operator exchanges genes between two parent individuals in order to create new child phenotypes. The pre-specified probability c is used to select a proportion of individuals for crossover. For example:

Parent 1: $\left(x_1 \quad y_1 \quad a_1 \quad b_1 \quad \theta_1 \quad Ts_1 \quad Tin_1\right)$

Parent 2: $\left(x_2 \quad y_2 \quad a_2 \quad b_2 \quad \theta_2 \quad Ts_2 \quad Tin_2\right)$.

If the crossover point is 6, the result after crossover will be (Figure 1):

Child1=$\left(x_1 \quad y_1 \quad a_1 \quad b_1 \quad \theta_1 \quad Ts_2 \quad Tin_1\right)$

Child2=$\left(x_2 \quad y_2 \quad a_2 \quad b_2 \quad \theta_2 \quad Ts_1 \quad Tin_2\right)$.

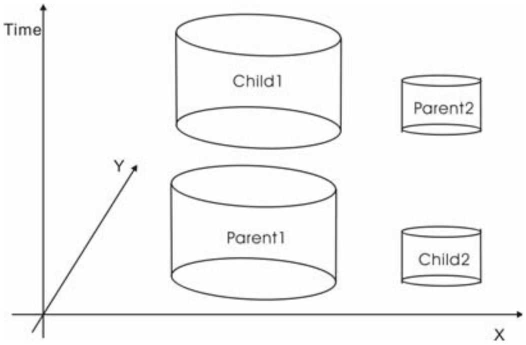

Figure 1 graphically illustrates a crossover example. The two parents exchange the 6th variables.

2.6 Mutation

Mutation is used to make a random change to an individual to increase or maintain population diversity. In the proposed genetic algorithm, the mutation operation will change one of the 7 parameters. The example below indicates how a new individual is created by mutation (Figure 2):

Parent individual: $\left(x_1 \quad y_1 \quad a_1 \quad b_1 \quad \theta_1 \quad Ts_1 \quad Tin_1\right)$

Randomly choose 1st parameter as the mutation point. The mutation operation generates a new individual as:

New individual: $\begin{pmatrix} x_1{}' & y_1 & a_1 & b_1 & \theta_1 & Ts_1 & Tin_1 \end{pmatrix}$.

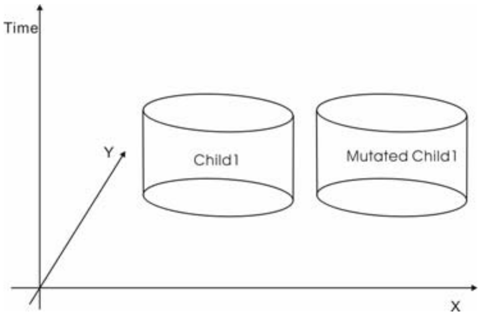

Figure 2 graphically illustrates a mutation example: the mutation point is on position 1.

2.7 Termination

The proposed algorithm is terminated after a pre-specified number of generations. During the previous generation to the next, if the fitness of an individual is larger than 0, which means the observed number of cases is larger than what is expected based on its population size, this individual will be exported as a potential cluster.

3. Application

Experimental data was obtained from http://www.satscan.org/datasets/. The real world dataset shows the spatial distribution of brain cancer incidence in New Mexico. The dataset has spatiotemporal attributes and is appropriate for testing the proposed genetic algorithm. The proposed GA is coded in MatLab 7.1 within Chipperfiled's GA toolbox. It runs on Dell workstation with Xeon^{TM} 2.66 GHz CPU and 2.00 GB RAM. Figure 3 shows two clusters of brain cancers. The time period of the first cluster is between 1985 and 1989. The second cluster has a time period of 1988–1989. Results are consistent with findings from Kulldorff's SatScan.

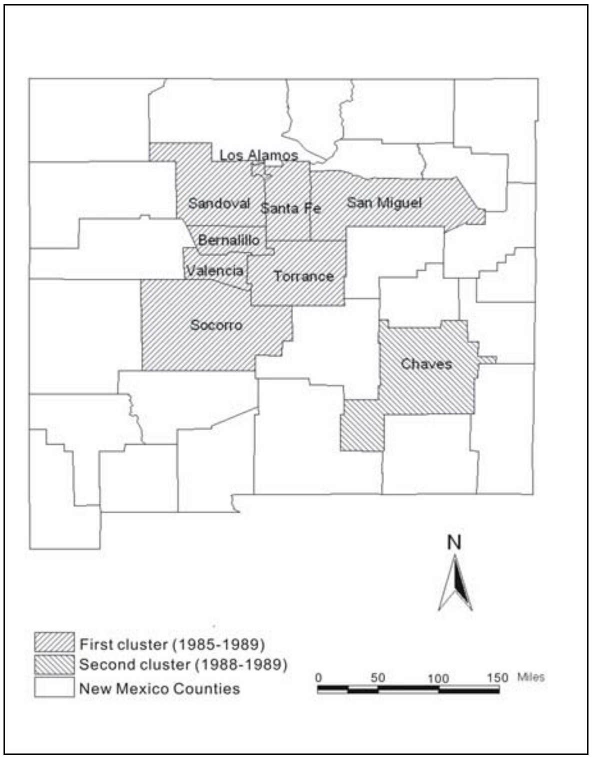

Figure 3 shows two spatiotemporal clusters of brain cancer incidence in New Mexico from 1973–1991.

4. Discussion and Ongoing Research

A GA-based approach has been presented to detect spatiotemporal clusters. The algorithm is suited to identifying clusters of rates/risks in relation to underlying populations at risk when both space-time and space-time-attribute are considered. The application of the approach shows that this method has great potential for spatiotemporal cluster detection.

In ongoing research, we are addressing the following issues:

1) How will a parameter configuration impact the performance of this genetic algorithm? How should a configuration be determined when the clusters are previously unknown?
2) How does this method perform with different datasets? Can it deliver runtime saving that a machine learning method can offer and at the same time not sacrificing cluster accuracy?
3) How does this GA perform compared against other methods? Under what condition will this genetic approach outperform others, and vice versa? How can this GA be extended to account for irregularly shaped clusters?

5. References

Conley J, Gahegan M, and Macgill J, 2005, A genetic approach to detecting clusters in point data sets. *Geographical Analysis*, 37:286-314.

Ester M, Kregel HP, Sander J, and Xu X, 1996, A density-based algorithm for discovering clusters in large spatial databases. In: *Proceedings of 2nd Int. Conf. on Knowledge Discovery and Data Minin*g, Portland, OR, 226-331.

Estivill-Castro V and Lee I, 2000, AUTOCLUST: Automatic Clustering via Boundary Extraction for Mining Massive Point-Data Sets. *Proceedings of the 5th International Conference on GeoComputation*.

Goldberg DE, 1989, *Genetic algorithms in search, optimization, and machine learning*, New York, Addision-Wesley.

Hobbs M and Goodchild MF, 1996, Spatial clustering with a genetic algorithm. In: Parker D, *Innovations in GIS 3*, London, Taylor & Francis, 85-93.

Karypis G, Han E, and Kumar V, 1999, CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *IEEE Computer: Special Issue on Data Analysis and Mining*, 32(8):68-75.

Kulldorff M, 2005, SatScan^{TM} User Guide. Available: http://www.satscan.org [accessed 27 November 2005].

Mantel M, 1967, The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27:209-220.

Openshaw S and Perrée T, 1996, User-centered intelligent spatial analysis off point data. In: Parker D, *Innovations in GIS 3*, London, Taylor & Francis, 119-134.

Openshaw S, Charlton M, Wymer C, and Craft A, 1987, A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information System*, 1:335-358.

Snow J, 1855, *On the mode of communication of cholera*. Churchill Livingstone, London.