Polygon-Based Spatial Clustering[1]

J. Zhang, A. Samal and L.-K. Soh

Department of Computer Science & Engineering
University of Nebraska-Lincoln
Lincoln, NE 68699-0115, USA.
Tel: (402) 472-2217
FAX: (402) 472-7767
Email: {jzhang, samal, lksoh}@cse.unl.edu

Abstract

Clustering geographic data using traditional methods often result in clusters that look dispersed over the geographic space and poorly reflect any underlying spatial structure. We propose a *polygon*-based spatial clustering approach, which models a spatial object as a *polygon* with three groups of attributes: *general attributes*, *boundary attributes*, and *spatial events*. We have developed a generalized distance function as a combination of distance functions defined on each group of attributes. The effectiveness of the approach is tested using a hydrological application. Experimental results show that this approach can organize the data into meaningful categories.

1. Introduction

Cluster analysis encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories (Jain 1999). The attractiveness of cluster analysis is its ability to find categories or clusters directly from the given data. Many clustering approaches and algorithms have been developed and successfully applied to many applications. However, when a classical clustering technique, such as the k-means, is applied to geographically located data, e.g., watershed data in this research, without using the spatial information, the resulting partition has often a "chaotic" appearance over the geographic space, i.e., clusters look dispersed, and reflect only poorly any eventual underlying spatial structure. This is because classical clustering algorithms often make assumptions (e.g., independent, identical distributions) which violate Tobler's first law of geography: everything is related to everything else but nearby things are more related than distant things (Tobler 1979). In other words, the values of attributes of nearby spatial objects tend to systematically affect each other.

Spatial clustering is a process of grouping a set of spatial objects into clusters. For instance, spatial clustering is used to determine the "hot spots" in crime analysis and disease tracking. Hot spot analysis discovers unusually dense event clusters across time and space. Many criminal justice agencies are exploring the benefits provided by computer technologies to identify crime hot spots in order to take preventive strategies such as deploying saturation patrols in hot spot areas (Han 2001).

---

[1] This work was supported in part by NSF ITR Grant NSF 01-149.

The implicit assumption of spatial clustering is that patterns in space tend to be grouped rather than randomly located. However, the statistical significance of spatial clusters should be measured by testing the assumption in the data. Also, spatial data can be described in terms of certain explanatory terms and obstacles, which cannot be incorporated in classical high dimensional cluster techniques (Tung and Hou 2001). For such spatial clustering problems, some spatial clustering techniques have been developed. CLARANS for example, is a clustering algorithm based on randomized search for spatial clustering (Ng and Han 1994). Guo et al. (2002) suggest a spatial clustering method, which combines the properties of low dimensional spatial clustering and high dimensional cluster identification.

Spatial objects such as points, lines, or polygons are all represented by a set of points. For example, a polygon can be represented by its edges (vector representation) or by the points contained in its interior, e.g., the pixels of an object in a raster image (raster representation). *Topological relations* are based on the boundaries, interiors and complements of two related objects and are invariant under transformations, which are continuous, one-one, onto and whose inverse is continuous. Some example relations are: *A disjoint B, A meets B, A overlaps B, A contains B, A inside B,* etc.

Spatial clustering techniques adapted from traditional clustering approaches usually only apply to a set of point objects (feature vectors), although each point object may have many dimensions (attributes). For instance, in the cluster analysis of watersheds, one can incorporate the centroid of a watershed as an attribute representing its geospatial location, and then cluster watersheds based on the Euclidean distance between their centroids. However, this straightforward approach does not consider important spatial properties of watersheds such as boundary and neighborhood. Meanwhile, two watersheds separated by plateau should have a different inter-watershed distance than those separated by mountain in terms of the elevation and aspect value along the boundary. We have developed a new polygon-based spatial clustering approach by incorporating the polygon boundary and spatial events in each polygon into a spatial clustering algorithm. We have subsequently implemented and applied the approach to cluster hydrological watersheds in the state of Nebraska. Results show that adding spatial attributes into the clustering process improve the coherence of the clusters.

2. Problem Formulation

Given a set of polygons $P = \{P_1, P_2, \cdots P_n\}$ in which each polygon P_i has three sets of attributes derived from geospatial data,

$$P_i =< B_i, G_i, S_i > \qquad (1)$$

where

B_i is the boundary of the polygon P_i,

G_i is the set of general attributes of the polygon P_i,

S_i is the set of spatial events in the polygon P_i.

The clustering problem can be defined as: given a data set of *n* polygon (non-point) objects with these attributes, find a partitioning of it into groups (clusters) with respect to some similarity measure or distance metric. The new spatial clustering algorithm shall use a generalized distance, which is a

combination of the differences in the non-spatial attribute values and distances defined on spatial constraints.

3. Approach

To address the specific problem of clustering objects that are spatially referenced, we propose a new spatial modelling approach that represents spatial objects as polygon objects with spatial attributes and design a new polygon-based spatial clustering algorithm. Specifically, a spatial object is a polygon object that contains three groups of attributes: general attributes, boundary attributes, and spatial events. The attributes used in the application of watershed study are briefly described below.

General attributes of watersheds consist of a set of numerical non-spatial features that include measurement of time series data collected from stations such as surface water stream gauges, groundwater wells, and weather stations. They also contain various shape parameters, e.g., area of Minimum Bounding Box (MBR), aspect ratio (elongation), and orientation (angle).

To avoid the disturbance of different geographical and meteorological factors in event of each individual watershed, we do not use the original time series data of each watershed for the clustering process. Instead we calculate and make use of changing trends and relations of these original factors. For example, we measure the correlation between a surface water flow and the change of local weather as a feature of the watershed. In some watersheds, the surface flow is affected and determined totally by the weather (temperature, precipitation and evaporation), so the correlation value may be very high. But in other areas, the surface flow may be supplemented and adjusted by the ground water, so the surface water flow and the local weather may not be tightly correlated. Here the correlation of surface water and weather is a distinguishing feature of a watershed.

Spatial attributes of watersheds consist of polygon boundary and spatial events inside the polygon. Polygon boundary is specified by a set of vertices defined in some spatial coordinate frame. Spatial events in each polygon are categorized into point, line and area to represent events such as the lake coverage in a polygon, the linear relationship of two neighboring polygons, and the elevation change along the shared boundary between two polygons.

For spatial clustering, if the attributes correspond to point objects, natural measurements of similarities such as Euclidean distance exist. However, if the attributes correspond to polygon objects, the situation is more complicated. More specifically, the dissimilarity (or distance) between two polygon objects may be defined in many ways, some better than others. But more accurate distance measurements may require more effort to compute. In our case, a generalized distance function (GDF) is developed for the polygon-based spatial clustering algorithm to determine how dissimilar two polygon objects are given the differences in the attribute values.

The GDF is a combination of three distance functions defined on each of the three group of attributes discussed earlier. A modified Hausdorff distance is used to measure the distance between two polygons because polygon objects that represent watersheds do not intersect and no polygon object contains the other. Meanwhile, Euclidean distance between two vectors is applied to the general attributes. The distance on spatial events is the most complicated one. It is a weighted sum of all the distances defined on different spatial events, considering different configurations of

polygons such as polygons with shared event on the boundary and connected events.

Experimental results on the watershed data in Nebraska show the effectiveness and efficiency of our polygon-based spatial clustering in watershed analysis. This approach can be applied to cluster other type of spatial objects.

4. Implementation and Results

We use the watershed polygons for the state of Nebraska to evaluation the clustering algorithm. There are 69 watersheds inside Nebraska; with more than 800 different kinds of hydrologic observation stations, including surface water stations, ground water stations, weather stations, etc. Some stations contain daily data starting even from the late nineteenth century. General attributes of the watersheds include the correlations between surface water, ground water, and precipitation, their trends, the coordinates of the geographic center of the watershed, the shape parameters, average elevation, etc. The spatial events used in our analysis include the lakes and streams. See Zhang (2004) for details of the algorithms and implementation.

We compare the experimental results of the standard k-means and those of our spatial clustering algorithm. It illustrates spatial variations of the representative watersheds given different weights of spatial distance functions. As discussed earlier, we used the distance function:

$$Distance(P_1,P_2) = \alpha_1 D_g(P_1,P_2) + \alpha_2 D_b(P_1,P_2) + \alpha_3 D_s(P_1,P_2) \qquad (2)$$

where D_g, D_b, D_s represents distance on general attributes, boundary features, and spatial events respectively, and $\alpha_1 + \alpha_2 + \alpha_3 = 1$. The distance on spatial events is

$$D_s(P_1,P_2) = \beta_1 D_p(P_1,P_2) + \beta_2 D_l(P_1,P_2) + \beta_3 D_a(P_1,P_2) \qquad (3)$$

where D_p, D_l, D_a represents spatial distance on point event, line event, and areal event respectively, and $\beta_1 + \beta_2 + \beta_3 = 1$.

4.1. Comparison on the Results of Spatial and Non-spatial Clustering

We used the same weight $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$ for each distance components. We used modified Hausdorff distance on the watershed boundary features and used distances computed from point spatial event (i.e., lake) and line spatial event (i.e., river). The optimal number of clusters with validity value in each experiment is summarized in Table 1, in which SGW_G_B represents the combination of general attributes and boundary feature, SGW_G_S represents the combination of general attributes and spatial events, SGW_G_B_S represents the combination of general attributes, boundary feature, and spatial events. The results show that using spatial attributes as described here results in better clustering results. According to experts, results obtained using spatial events are more in line with their view of the hydrologic landscape.

Table 1. Optimal number of clusters with validity value

	Optimal Number of Clusters	Validity Index
SGW_G	4	1.2631
SGW_G_B	4	1.3608
SGW_G_S	4	2.1940
SGW_G_B_S	4	2.1158

5. Summary and Conclusions

We have developed a polygon-based spatial clustering algorithm for spatial objects. A spatial object is modeled as a polygon that contains three set of attributes: general attributes, boundary attributes and spatial events. The algorithm uses a generalized distance function defined on all these attributes of the polygon objects. The approach has been implemented and tested on natural resource data, i.e. watershed data, from the State of Nebraska. We have processed a large set of data for generating all the features needed in the clustering procedure. Three datasets with different features have been used and the experimental results reveal some interesting and useful patterns in the watersheds.

6. References

Jain, A.K, Murty, M.N., and Flynn P.J. 1999. Data clustering: a review. ACM Computing Surveys, 31, 3, 264-323.

Tobler, W.R. 1979. Cellular Geography, Philosophy in Geography. Gale and Olsson, Eds., Dordrecht, Reidel.

Han, J., Kamber, M., and Tung, A., 2001a, Spatial Clustering Methods in Data Mining: A Survey. In Miller, H., and Han, J., eds., Geographic Data Mining and Knowledge Discovery. Taylor and Francis. 21.

Tung, A. K. H., Hou, J., et al., 2001, Spatial clustering in the presence of obstacles. The 17th International Conference on Data Engineering, ICDE'01.

Ng, R. T., Han, J., 1994, Efficient and Effective Clustering Methods for Spatial Data Mining. 20th International Conference on Very Large Data Bases, September 12-15, Santiago.

Guo, D., Peuquet, D., and Gahegan, M., 2002, Interactive Subspace Clustering for Mining High-Dimensional Spatial Patterns. Second International Conference on Geographic Information Science, Boulder, Colorado, USA. Accessed at http://www.geovista.psu.edu/publications/10421.pdf

Zhang, J. 2004. Polygon-based Spatial clustering and its application in watershed study. MS Thesis, Department of Computer Science and Engineering, University of Nebraska-Lincoln, December 2004.