KNOWLEDGE DISCOVERY BY SPATIAL CLUSTERING BASED ON SELF-ORGANIZING FEATURE MAP AND A COMPOSITE DISTANCE MEASURE

Limin Jiao[a,b], Yaolin Liu*[a,b]

[a]School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan, P.R.C., 430079.
lmjiao027@163.com, - yaolin610@163.com
[b]Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University

WG II/2

KEY WORDS: Self-Organizing Feature Map, Spatial Clustering, Knowledge Discovery, Data Mining

ABSTRACT:

This paper proposes a spatial clustering model based on self-organizing feature map and a composite distance measure, and studies the knowledge discovery from spatial database of spatial objects with non-spatial attributes. This paper gives the structure, algorithm of the spatial clustering model based on SOFM. We put forward a composite clustering statistic, which is calculated by both geographical coordinates and non-spatial attributes of spatial objects, and revising the learning algorithm of self-organizing clustering model. The clustering model is unsupervised learning and self-organizing, no need to pre-determine all clustering centers, having little man-made influence, and shows more intelligent. The composite distance based spatial clustering can lead to more objective results, indicating inherent domain knowledge and rules. Taking urban land price samples as a case, we perform domain knowledge discovery by the spatial clustering model. First, we implement several spatial clustering for multi-purpose, and find series spatial classifications of points with non-spatial attributes from multi-angle of view. Then we detect spatial outliers by the self-organizing clustering result based on composite distance statistic along with some spatial analysis technologies. We also get a meaningful and useful spatial homogeneous area partition of the study area.

1. INTRODUCTION

Spatial data mining and knowledge discover (SDMKD) is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases (Roddick & Spiliopoulou1999; Li, 2001, 2002; Shekhar & Chawla, 2001, 2002). Types of the knowledge may be discovered from spatial database usually include spatial characterization, spatial classification, spatial dependency, spatial association, spatial clustering and spatial trend/outlier analysis.

Spatial clustering is the process of grouping a set of spatial objects into meaningful subclasses (that is, clusters) so that the members within a cluster are as similar as possible whereas the members of different clusters differ as much as possible from each other. Spatial clustering can be applied to group similar spatial objects together, and its implicit assumption is that patterns tend to be grouped in space than in a random pattern. The statistical significance of spatial clustering can be measured by testing the assumption in data. After the verification of the statistical significance of spatial clustering, and clustering algorithms are used to discover interesting clusters.

Clustering of points is a most typical task in spatial clustering, and many kinds of spatial objects clustering can be abstracted as or transformed to points clustering. This paper will discuss this kind of spatial clustering. In many situations, spatial objects are represented by points, such as cities distributing in a region, facilities in a city, and spatial sampling points for some research reason, e.g. ore grade samples in ore quality analysis, elevation samples in terrain analysis, land price samples in land evaluation, etc. The analysis and data mining of spatial points include spatial distribution characters calculation (e.g. density, centers, and axes), spatial trend analysis, spatial clustering, etc. Spatial clustering is the most complicated in these. Clustering of spatial points is to divide a group of point objects into some more inner-accumulative sub-groups according to given regulations.

Based on the technique adopted to define clusters, the clustering algorithms can be divided into four broad categories(Han, Kamber, & Tung, 2001; Halkidi, Batistakis, & Vazirgiannis, 2001; Chawla & Shekhar, 2001): Hierarchical clustering methods (AGNES, BIRCH[10], etc.), Partitional clustering algorithms(K-means, K-medoids, etc.), Density-based clustering algorithms(DBSCAN、DENCLUE[12], etc.), Grid-based clustering algorithms (STING[15], etc.). Many of these can be adapted to or are specially tailored for spatial data (Han, Tung and Kamber, 2001). Theoretically, these methods of general clustering can also be used in spatial clustering when we add (x, y) coordinates into the input vector as other variables. But it is not proper because these methods are not developed specially for spatial clustering.

Important data characteristics, which affect the efficiency of clustering techniques, include the type of data (nominal, ordinal, numeric), data dimensionality (since some techniques perform better in low-dimensional spaces) and error (since some techniques are sensitive to noise) (Han, Kamber and Tung

* Corresponding author: Liu Yaolin, E-mail:yaolin610@163.com; This paper is funded by 863 program of China (2007AA122200), 973 program of China (2006CB701303) and Open Research Fund Program of GIS Laboratory of Wuhan University（wd200609）

2001). In this paper, the spatial clustering method based on self-organizing neural networks is for numeric data, and is suitable for both low-dimensional and high-dimensional data spaces, even with noise.

Many approaches in spatial clustering focus on pure geometric clustering, i.e., spatial clustering is only based on geometric distances between objects or other measure derived from location (e.g. dispersion). These methods over-simplify the spatial clustering in some sense, and cannot perform better in clustering of spatial points with non-spatial thematic attributes. In advance clustering analysis and data mining of spatial points, thematic attributes and locations are to be considered compositively. The basic statistic parameter in clustering, the distance between spatial points, is to be defined properly. Zhang (2007) presented a Penalized Spatial Distance (PSD) measure to guarantee the non-overlapping geographic constraint in spatial clustering with non-spatial attributes. This is an achievable attempt that adjusting spatial distance by using penalizing factor. In this paper, we will propose a composite distance measure considered both location and non-spatial attributes.

This paper puts forward a clustering method of spatial points based on self-organizing feature map (SOFM) and a composite distance measure. Further more, this paper discusses data mining and knowledge discovery of spatial points based on self-organizing clustering results, including spatial classification, detection of spatial outliers and homogeneous area partition.

2. SPATIAL CLUSTERING BASED ON SELF-ORGANIZING FEATURE MAP

2.1 Self-organizing Feature Map (SOFM)

Self-organizing feature map (SOFM), proposed by Prof. Kohonen, is a kind of self-organizing neural network. In human brain, adjacent neurons activate each other, and distant neurons restrain each other, but a faint activation will be found between more distant neurons. SOFM imitates this kind of interaction by competitive learning. In the network training, neurons compete with each other, and the winner will activate neighboring neurons and restrain the others. Along with the training, the network produces some self-organizing output sub-populations, which represent different samples distribution. SOFM can map input data to 2-dimentional space (plane). And it can divide spatial objects into sub sets according to their inter-distances, so can be used in spatial clustering[5, 7].

The structure of SOFM is shown in Fig.1. The input neurons connect to each output neuron. Sometime there is connection among the neurons in compete layer (output layer), which control the interaction among output neurons. It also works when employ neighbor activating but not connections among output neurons, and the algorithm becomes brief. So do we in this paper.

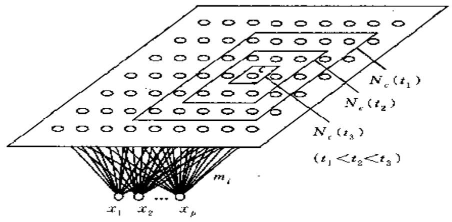


Figure 1. The Structure of SOFM and the neighborhood of competitive neuron

SOFM maps the input data onto the competitive layer with M neurons, a 2-dimensional array, which represents the classification mode, and each output node (neuron) indicates a cluster after training. Associated with each node *j* is a parametric reference vector $m_j = \{m_j^{(d)}, d = 1,2,...,D; j = 1,...,M\}$, where $m_j^{(d)}$ is the connection weight between node *j* and input *d*, D is the dimension of input vector, and M is the number of output nodes. The input data set, $x = x_n^{(d)}(n = 1,2,...,N; d = 1,2,...,D)$, consisting of *d*-dimensional input vector $x_n^{(d)}$, can be visualized as being connected to all nodes in parallel via a kind of scalar weights $m_j^{(d)}$. The aim of learning is to map all the *N* input vectors $x_n^{(d)}$ onto corresponding reference vector by adjusting weights $m_j^{(d)}$ such that the SOM gives the best match response locations. Therefore, the learning algorithm can be given as fowling:

(1) Initialize each connection weight to a random real number in [0,1];

(2) Select a sample x_n (input vector) random in samples set to feed the network;

(3) The input vector x_n is compared to all reference vectors $m_j^{(d)}$, searching for the smallest Euclidean distance to find the best matching node, signified by *c*.

$$\|x_n - m_c(t)\| = \min_j \{\|x_n - m_j(t)\|\} \tag{1}$$

where t is discrete time coordinate, $\|x_n - m_j(t)\|$ represents 2-norm (Euclid norm), i.e.

$$\|x_n - m_j(t)\| = \sqrt{\sum_{d=1}^{D}(x_n^{(d)} - m_j^{(d)}(t))^2} \tag{2}$$

(4) During the learning process the node that best matches the input vector is allowed to learn. Those nodes that are close to the node up to a certain distance will also be allowed to learn. The learning process is expressed as:

$$m_j(t+1) = \begin{cases} m_j(t) + \mu(t)h(t)[x_n - m_j(t)],\ j \in N_c \\ m_j(t),\ j \notin N_c \end{cases} \quad \textbf{(3)}$$

Where $\mu(t)$ is learning step, $h(t)$ is the size of the neighborhood of node *c*, signified by N_c .

(5) if $n < N$, then $n \leftarrow n+1$, go to (2);

(6) $t \leftarrow t+1$, go to (2) until the network has converged.

To make the model converge faster, the learning step and the size of neighborhood will decrease along with circulation. After the 2-dimensional map has been trained, each input vector is mapped onto the best matching node, the center of the cluster of corresponding input vectors.

2.2 The Structure of the Spatial Clustering Model Based on SOFM

The input vector of the SOFM model used in spatial clustering consists of geometric coordinates and thematic attributes of spatial points. For discrete points $P_n(n = 1,2,..., N)$, we can express their geometric locations as (x_n, y_n) , and their attributes, $z_n = (z_{n1}, z_{n2},..., z_{nm})$. All sample points construct the input set of SOFM, defined as

$$x = \{x_n \mid n = 1,2,...,N\} \quad \textbf{(4)}$$

$$x_n = \{x_n^{(1)}, x_n^{(2)}, x_n^{(3)},..., x_n^{(D)}\} = \{x_n, y_n, z_{n1},..., z_{nm}\} \quad \textbf{(5)}$$

Where, x_n represents the sample n (a spatial point), $x_n^{(1)}, x_n^{(2)}, x_n^{(3)},..., x_n^{(D)}$ are variables of sample n, $x_n, y_n, z_{n1},..., z_{nm}$ are coordinate x, coordinate y and m attributes of sample n. N represents the number of points, and D represents the dimension of the input vector. Obviously, D = m + 2, and the number of input neurons is equal to the dimension of the input vector.

The number of output neurons is set to the expectant number of clusters. The reference vectors are initialized random as original centers of clusters before training. Some clusters their distance is small enough can be merged if necessary after training.

To improve clustering result and accelerate training, it is necessary to make all variables of input vector in similar or the same range. So we need to standardize the input data. A variable is standardized by

$$x_i = \frac{x_i^{'} - x_{\min}}{x_{\max} - x_{\min}} \quad \textbf{(6)}$$

Where $x_i^{'}$ is the value of variable *i* before standardization, and x_i is the one after standardization, $x_{\min}$ is the minimum value of the variable, and $x_{\max}$ is the maximum.

3. COMPOSITE DISTANCE MEASURE

3.1 Definition of Composite distance measure

There are two kinds of clustering statistics usually used in spatial clustering. One is “distance”, which requires that the distance between any two samples in the same subpopulation is smaller then in different subpopulations. And the other is variance, which makes the variance (mean square) in every subpopulation as small as possible. The calculation of “distance”, the basic statistic, is the sticking point. In many applications, the spatial objects usually not only have geometric features (location), but also have some non-spatial thematic attributes. So the “distance” in spatial clustering must maintain these two aspects, spatial location and non-spatial attributes.

The “distance” in spatial clustering, especially in multidimensional clustering, should be defined as a kind of generalized distance, which consists of geometric coordinates and thematic attributes, and measures the neighborhood, similarity and relativity between spatial points with thematic attributes. In this paper, a composite distance is defined as

$$D_{ij} = w_p \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} + w_a \sqrt{\sum_{k=1}^{m} w_k (z_{ik} - z_{jk})^2}$$

$$\text{and } w_p + w_a = 1, \ \sum_{k=1}^{m} w_k = 1 \quad \textbf{(7)}$$

Where, D_{ij} is the composite distance between sample *i* and sample *j*, (x_i, y_i) and (x_j, y_j) are the standardized values of the coordinates of sample *i* and sample j, z_{ik} and z_{jk} represent the standardized values of attribute *k* of sample *i* and sample *j*, *m* is the number of thematic attribute of samples, w_p and w_a represents the weights of geometric distance and non-spatial attributes distance, w_k is the weight of attribute *k*. where w_p + w_a =1, and $\sum_k w_k = 1$. When $w_p = 1$, $w_a = 0$, D_{ij} is equal to geometric distance, the spatial clustering becomes geometric clustering. When $w_p = 0$, $w_a = 1$, D_{ij} is the distance of non-spatial attributes, the spatial clustering becomes semantic clustering.

3.2 Composite distance based spatial clustering

The training algorithm of SOFM used in spatial clustering should be revised according to the definition of generalized Euclidean distance. The input vector is also same to 2.2, and the step (3) of the training algorithm in 2.1 is revised as fowling:

The input vector x_n is compared to all reference vectors $m_j^{(d)}$, searching for the smallest Euclidean distance to find the best matching node, signified by *c*.

$$D_{nc}(t) = \min_j \{D_{nj}(t)\} \qquad (8)$$

Where, $D_{nc}(t)$ is the generalized Euclidean distance between input node *n* and output node *c*, and $D_{nj}(t)$ is the one between input node *n* and output node *j*, *t* is discrete time coordinate.

4. KNOWLEDGE DISCOVERY BY SELF-ORGANIZING SPATIAL CLUSTERING

4.1 Spatial Data

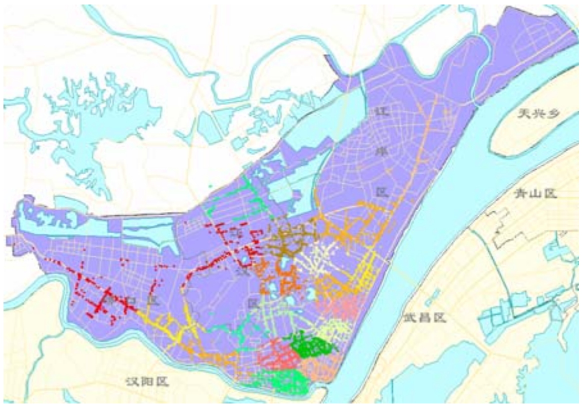
Here we take the land price samples in datum land price evaluation of Wuhan city, P.R.China, as a case to study the efficiency and application of the method proposed in this paper. Land price samples have three basic variables, coordinate *x*, coordinate *y*, and land price *p* that represents the price of per square meter land. For convenience, land price samples for commercial use in Hankou town, main commercial district in Wuhan, are used in this experiment. There are total 5967 samples, represented by spatial points. With the help of clustering technologies and some spatial analysis methods, we can mine knowledge implicated in spatial points.

4.2 Spatial Classification

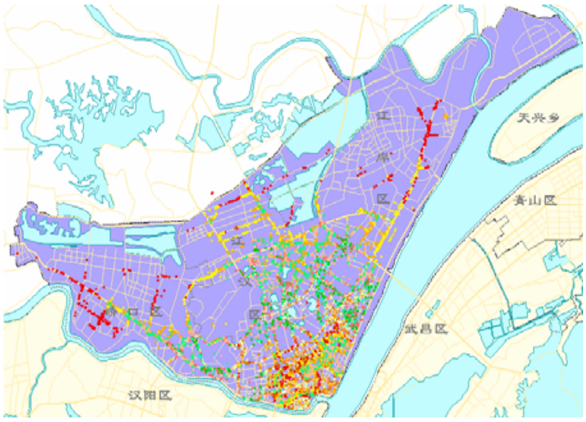
We know that we will get different clustering results by employ different clustering methods or distinct distance measures. Each clustering result indicates a classification with special meaning of original data set.

We can classify spatial points flexibly by self-organizing clustering. We can implement geometric clustering, non-spatial attributes clustering and multidimensional clustering by adjusting the weights of geometric distance and semantic distance in the composite distance formula. The SOFM used in spatial clustering of land price samples has 3 neurons in input layer, and the input vector is (x, y, p). The coordinates and land value are standardized according to the method in 2.2. After SOFM training, we get the result of self-organizing spatial clustering of land price samples.

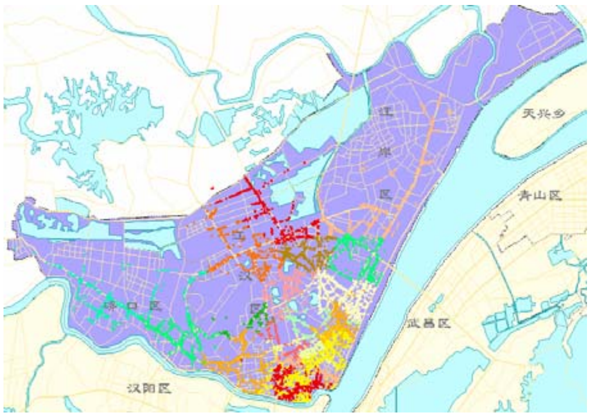
For comparison, we firstly use geometric distance and non-spatial distance to get different clustering results, see Fig.2(a) and Fig.2(b). Then we employ the composite distance measure given in 3.1 to calculate the adjacency between input sample and reference vectors, the result is shown in Fig.2(c).



(a) Geometric distance based spatial clustering



(b) Non-spatial attributes distance based spatial clustering



(c) Composite distance based spatial clustering

Figure 2. The result of self-organizing clustering of land price samples

Geometric clustering result indicates the characteristics of spatial distribution of points, which determined only by geometric distances among points. Non-spatial attributes based clusters are classified by thematic attributes, in this case, land price. We cannot see clear rules in Fig.2(b).

Usually, multidimensional clustering shows the hidden congregative characters of spatial points more objectively because it considers both geometric adjacency and semantic similarity. For land price samples, the clustering based on semantic attribute cannot resist the influence of random error, and geometric clustering may generate subpopulations with thematic value-diverse samples. We can see from Fig.2(c) that the subpopulations are spatially continuous to some extent, but geometric location is not the only determinant. Subpopulations seem to hint that where they locate are homogeneous areas.

Further more, in the interpolation of basic land price, we usually use neighboring samples to evaluate the value of unknown point, because neighboring land samples influence each other and tend to have similar unit price. But adjacency is not the only reason of the convergence of land unit price, so homogenous land price areas are not all circular, but in many kinds of shapes, e.g., zonal shape. It is necessary that searching neighbor samples in a homogenous area when estimate the value of an unknown point.

In a word, we can implement spatial clustering for multi-purpose using SOFM and composite distance measure, and can find series spatial classifications of points with non-spatial attributes from multi-angle of view.

4.3 Spatial Outlier Detection

Ordinary, we have following hypothesis: If there are a very few samples distributing in the scope of a samples set, whose members belong to the same cluster and their domains are spatial continuous, then these samples are regarded as spatial outliers, which may be the samples with gross error or the spatial outliers with special meaning. We can detect spatial outliers based on the self-organizing clustering result along with some spatial analysis technologies.

We employ Voronoi diagram to analyze the distributing character of clusters. The Voronoi diagram of a set of n points on a plane divides the plane into n polygons, and any location in each polygon is closer to the point contained in this polygon than any other points. We produce the Voronoi diagram of land price samples, and use the outline of the case area as the boundary constraint. Signify all Voronoi polygons with different colors according to clustering result (Fig.3). Aggregate the polygons those belong to the same class (Fig.4), and pick out the points within the tiny and fragmentary polygons as spatial outliers (Fig.5).

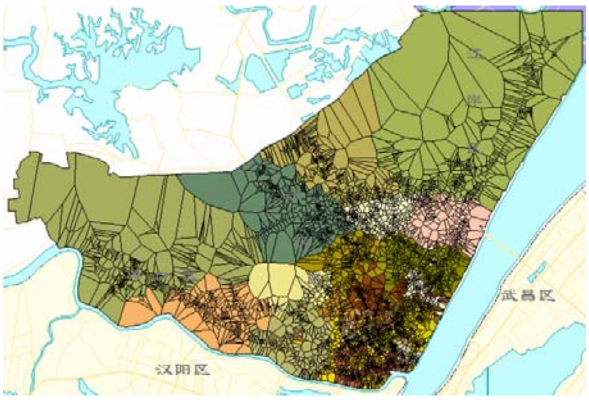


Figure 3. Voronoi polygons of land price samples (Signified according to clustering result)

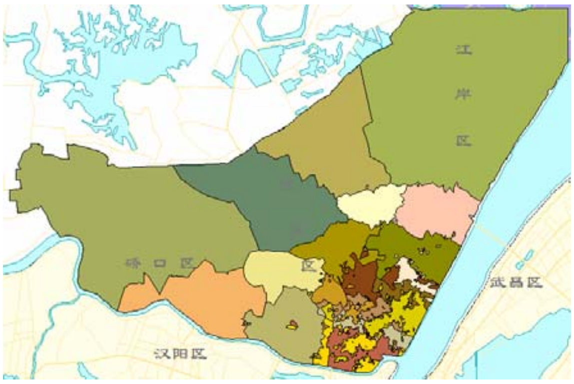


Figure 4. Aggregation of Voronoi polygons belong to the same class

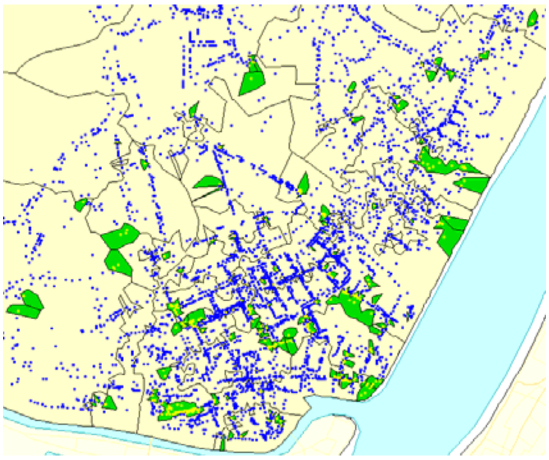


Figure 5. Fragmentary polygons and spatial outliers

We will find that these spatial outliers can be divided into two classes by examining them via additional profiles or checking on the spot. The major of them are samples with gross errors, which may be caused by data acquisition or data process. The others are special samples with the land price according with market condition, which hint that there are some particular local factors affecting land price in those locations, such as psycological reasons, local conditions, etc..

4.4 Homogeneous Area Partition

After the scattered small polygons was aggregated to the polygons those have longest sharing boundary with them, we get a partition of the case area, see Fig.6.

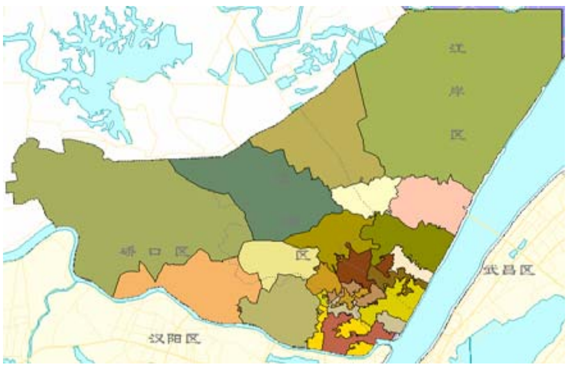


Figure. 6. Amalgamation of fragmentary polygons (homogenous partition)

This partition considers both geometric and thematic distances among spatial points, and avoids the influence of the samples with gross errors. Sub regions represent homogenous land price zones. We get more meaningful spatial partition knowledge by SOFM, spatial analysis technology and composite distance measure. This partition can be used in land grading (grading based on land price samples), or as a reference and comparison of the grading result generated by other methods. Homogeneous area partition reduces the complexity of the whole data space, and also can be used in local interpolation of basic land price.

5. CONCLUSIONS AND FUTURE WORK

This paper proposes a spatial clustering model based on SOFM and a composite distance measure, and studies the knowledge discovery by this self-organizing spatial clustering model. We have following conclusions:

(1) The spatial clustering based on SOFM is unsupervised learning and self-organizing, and no need to pre-determine all clustering centers. It has very little man-made influence, and shows more intelligent.

(2) The composite distance measure, used as clustering statistic, consists of both geometric distances and non-spatial thematic attributes, and can lead to more objective results, indicating inherent domain knowledge and rules. The clustering method using composite distance is more flexible, and can implement many kinds of spatial clustering for different aims.

(3) We can perform domain knowledge discovery by the spatial clustering model in this paper. First, we can implement spatial clustering for multi-purpose, and can find series spatial classifications of points with non-spatial attributes from multi-angle of view. Then we can detect spatial outliers by the self-organizing clustering result based on composite distance statistic along with some spatial analysis technologies. We also can get meaningful and useful spatial homogeneous area partition.

There are some issues to be studied further, such as knowledge discovery from series spatial clustering results when the weights in composite distance formula changes continuously, the application of self-organizing spatial clustering to other types of spatial objects, etc.

REFERENCES

Roddick, J.-F., and Spiliopoulou, M, 1999. A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research. SIGKDD Explorations, 1(1), pp. 34-38.

Shekhar, S., Schrater, P. R., et al., 2002. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. IEEE Transaction on Multimedia. 4(2).

Sanjay Chawla, Shashi Shekhar, et al., 2003. Modeling Spatial Dependencies for Mining Geospatial Data: An Introduction. Geographic Data Mining and Knowledge discovery (GKD). http://www.geog.utah.edu/ ~hmiller/gkd text.

Han, J., Kamber, M. and Tung, A. K. H. (2001) "Spatial clustering methods in data mining: A survey," in H. J. Miller and J. Han (eds.) Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, pp. 188-217.

Zhang Naiyao, Yan Pingfan, 1998. Neural networks and fuzzy control. Beijing: Tsinghua University Press (Chinese)

Guo Renhong, 2001. Spatial analysis, Beijing: Higher Education Press (Chinese)

Kohonen, T, 1990. The Self-organizing Map, Proceedings of the IEEE, 78(9), pp. 1464-1480.

Pal N R, et al., 1993. Generalized Clustering Networks and kohonen's Self-organizing Scheme, IEEE, Trans NN, 4, pp. 549~557

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

T. Zhang, R. Ramakrishnan, M. Livny, 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases,Proc. 1996 ACM- SIGMOD Int. Conf. Management of Data (SIGMOD'96), pp. 103-114

G. Karypis, E.-H. Han, V. Kumar, 1999. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling, COMPUTER, 32, pp. 68-75

J. Sander, M. Ester, H.-P. Kriegel, X. Xu, 1998. Density-based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications, Data Mining and Knowledge Discovery, 2, 2, pp. 169-194

M. Ester, H. P. Kriegel, J. Sander, 2001. Algorithms and Applications for Spatial Data Mining, Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis

A. Hinneburg, D. A. Keim, 1998. An Efficient Approach to Clustering in Large Multimedia Databases with Noise, Proc. 1998, Int. Conf. Knowledge Discovery and Data Mining (KDD'98), pp. 58-65

J. Wang, R. Yang, Muntz, 1997. STING: A Statistical Information Grid Approach to Spatial Data Mining, Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97), pp. 186-195

Bin Zhang, Wen Jun Yin, Ming Xie, and Jin Dong, 2007. Geo-spatial Clustering with Non-spatial Attributes and Geographic Non-overlapping Constraint: A Penalized Spatial Distance Measure, PAKDD'07, Nanjing, China.