Mining Knowledge from Result Comparison Between Spatial Clustering Themes

SHA Zongyao  BIAN Fuling

ABSTRACT  This paper introduces some definitions and defines a set of calculating indexes to facilitate the research, and then presents an algorithm to complete the spatial clustering result comparison between different clustering themes. The research shows that some valuable spatial correlation patterns can be further found from the clustering result comparison with multi-themes, based on traditional spatial clustering as the first step. Those patterns can tell us what relations those themes have, and thus will help us have a deeper understanding of the studied spatial entities. An example is also given to demonstrate the principle and process of the method.

KEYWORDS  GIS; knowledge mining; spatial clustering; themes; spatial information representation; algorithms

CLC NUMBER  P208

Introduction

Knowledge discovery in databases (KDD) and data mining (DM) have been an area of increasing interests during recent years. Because data mining can extract desirable knowledge or interesting patterns from existent databases and ease the development bottleneck in building expert systems so they have become common interest to researchers in machine learning, pattern recognition, statistics, artificial intelligence, and high performance computing[13]. According to knowledge’s types, it can be classified as spatial information generalization, spatial association, spatial classification and spatial clustering, etc. Spatial clustering analysis subdivides all to-be clustered samples into several subgroups according to the principle of “minimum homogeneity within group and maximum heterogeneity between groups” and has potential value in spatial analysis and decision-making as region planning, soil classification and environmental quality evaluation. The current research on spatial clustering, however, only gives the possible aggregation character of clustered samples, which cannot represent the relation between samples clustered in different groups. It is necessary to further the analysis between those samples. Spatial entity (referred as spatial clustering samples here) is composed of various spatial dimensions in certain spatial information representation mode. Those spatial dimensions pave the result comparison between different spatial clustering themes. A spatial clustering theme is composed of several united dimensions according to application objective. So different application objectives may have different clustering theme. It is known that those themes are not independent of each other and should be associated together to give a complete expression of spatial objects. Some patterns can be mined by comparison of those clustering results. We present in this paper principles and approaches to analyze the

Received on December 12, 2004.
SHA Zongyao, Ph.D., School of International Software, Wuhan University, 129 Luoyu Road, Wuhan 430079, China, E-mail: zongyaosha@163.com
clustering results between different themes.

1 Basic concepts

1.1 Spatial information representation

Spatial information representation (or SIR) abstracts and redraws spatial entities of the real world according to certain concepts, rules or principles extracted from the human comprehensive understanding of the world, thus offers a basic tool to transmit ideas and thoughts among persons and objects. GIS abstracts spatial entities into various objects under the general principle of SIR to fit for automatic computer processing. So SIR is a requisite to realize spatial analysis. The general principle of SIR for GIS requires that a spatial entity can be expressed by various dimensions, with each describing one aspect of the entity. For example, a spatial entity may be constructed with an attribute set composed of n-dimensions.

1.2 Spatial clustering analysis

Clustering is to divide a series of samples into several meaningful sub-groups that samples in each sub-group are similar in character while samples in different sub-groups are distinctive. Clustering analysis is "a multivariate statistical procedure that starts with a data set containing information about a sample of entities and attempts to organize these into relatively homogenous groups"[4]. Cluster analysis often uses a dendrogram, which arranges multivariate data sets hierarchically according to their similarity[4].

2 Spatial clustering themes and comparison

2.1 Sampling and granularity of spatial clustering

Spatial clustering analysis classifies and aggregates spatial entities to uncertain groups based on the relation and difference of those entities. Those entities are individual element of cluster and are defined as clustering samples statistically. The whole set of samples is therefore defined as spatial clustering space in which spatial clustering is executed. Granularity of spatial clustering determines the scale of clustering and the clustered number of groups. Administrative regions of China can be clustered on the level of province or city. The higher the granularity is, the smaller the clustered number of groups will be achieved.

2.2 Dimensions and themes of spatial clustering

According to the spatial information classification of SIR, it is necessary to analyze a spatial entity from different aspects in order to understand it. Similar to describe spatial location with coordinate $x/y$ in a plane geometry or $x/y/z$ in a solid geometry, a spatial entity is an object in multi-dimensional space. A cluster dimension of field in spatial clustering corresponds with a description aspect of a spatial entity. A spatial clustering theme is composed of one or composite cluster dimensions. For example, if we take a spatial coordinate as a dependent theme, it can be plane coordinate composed of $x/y$ or solid coordinate composed of $x/y/z$. Spatial clustering can take different clustering themes. Fig. 1(a), Fig. 1(b) and Fig. 1(c) are the spatially clustered results of soil samples collected randomly on the basis of spatial theme (two dimensions), time theme (one dimension) and attribute theme (three dimensions). What theme should be set is determined by our concrete application. Sometimes clustering themes can also be united further to form higher theme. This paper will discuss spatial clustering of geometric theme (as Fig. 1 (a) shows) and attribute theme (as Fig. 1(c) shows). Then the clustering result comparison will be researched further.

2.3 Correlative index and inclusive index between spatial clustering themes

Correlative and inclusive analysis can be made between spatial clustering themes, based on
clustering samples as band. Correlative and inclusive index reflects an intrinsic relation between analyzed themes and can be used to make quantitative analysis. We take samples as shown in Fig. 1, which can be clustered into three types according to the attribute theme and the same is true to the geometric theme although the samples contained in each sub-group are not parallel. This indicates that there exists some correlation between the two themes even the correlation relation may be not true to all the themes. It may be a natural rule (in this sense it is also knowledge) implicit to us before it is found out.

By further comparing the content of sub-groups based on the two themes, it is found that a certain percentage of samples are the same between the two clustering themes. Statistically, it is regarded that the attribute of the studied samples may be dependent on its spatial location and distribution somewhat thus the attribute and location are not independent of each other.

There does not exist an effective measure to quantitatively represent the correlation between clustering themes. According to the statistics and set theory, the inclusion and relation of spatial samples will be adopted here to correlative index and inclusive index between spatial clustering themes.

As for Group A and Group B (A and B are also used to stand for their sample sets), we define:

- \( A \cap B : A \cup B \) is the inclusive index for Group A and Group B.

The following definition can be induced.

**Non relative:** If \( A \cap B : A \cup B = 0 \), then A and B has no relation (as shown in Fig. 2(a)), signed as \( A = 0 \).

**Partially relative:** If \( 0 < A \cap B : A \cup B < 1 \), then A and B are partially related (as shown in Fig. 2(b)), signed as \( A = k \), where \( 0 < k < 1 \).

**Full relative:** if \( A \cap B : A \cup B = 1 \), then A and B are full related (as shown in Fig. 2(c)), signed as \( A = 1 \).

**Non inclusive:** If \( A \cap B / B = 0 \), then B is not included in A or A does not include B, signed as \( B = 0 \).

**Partially inclusive:** If \( 0 < A \cap B : A \cup B < 1 \), then B is partially included in A or A partially includes B, signed as \( B = k \), where \( 0 < k < 1 \).

**Full inclusive:** If \( A \cap B / B = 1 \), then B is fully included in A or A completely includes B, signed as \( B = 1 \).

The correlative index and the inclusive index are the two important indexes to realize association analysis between clustering themes and they have the following relations:

- It is always true for \( B : A \leq B \ A \).
- If \( B : A \neq 1 \), then \( B : A = 0 = B \ A = 0 \) or \( B : A = 0 = B \ A = B \ A \).
- \( B : A = 1 \) and \( A = 1 = B \ A = 1 \).

The correlative index and the inclusive index can be used independently to make association analysis between clustering themes in applications. But they fit for different circumstances.
Generally speaking, if the analyzed themes are parallel, the relative index is preferred because it can unveil relevant or associative relation between the two themes. If the two clustering themes have principal and subordinate relations, the inclusive index is better because it is easy to get the reliable degree of subordinate theme to the principal theme. Take the full inclusive relation of Fig. 2(d) as example. If A is a sub-group of clustered result on the attribute theme and B is the other sub-group of clustered result on the geometric theme, then the inclusive index can be used to get the reliable degree of attribute to location. If the relative index is adopted it will be easy to find the cross relation between the two themes.

2.4 Similarity index and similarity matrix

The clustering analysis is usually based on a similarity matrix, so it is necessary to construct a similarity matrix. In multi themes clustering analysis, each theme is composed of several clustering dimensions. The most important step in the construction of the matrix is analyzing those dimensions and forms a weight matrix for them. In order to erase the influence of scale difference of those dimensions, the normalization transformation is required to make every dimension fit for the normal distribution signed as \( N(1,0) \) distribution (the standard error is 1 and average value is 0 for each dimension).

Let the dimension number of Theme \( i \) be \( n \), then the weight matrix for those dimensions is expressed as:

\[
Q_i = \begin{bmatrix}
q_{i1} \\
q_{i2} \\
\vdots \\
q_{in}
\end{bmatrix}
\]

Each factor in weight matrix \( q_{ij} (j = 1, 2, \ldots, n) \) stands for the weight value of Dimension \( j \) of Theme \( I \), where \( \sum q_{ij} = 1 \) \( (j = 1, 2, \ldots, n) \). The domain expert will give a weight value if no efficient data can be offered to make main factor statistical analysis.

The similarity matrix of Theme \( i \) is composed of dimensions and expressed as: \( W_i = \{ w_{i1}, w_{i2}, \ldots, w_{in} \} \). The calculated value for each element in the similarity matrix is: \( a_{ik} = W_i \times Q_i \).

Suppose the clustered sample number is \( m \) then the similarity matrix will be:

\[
A_{mn} = \begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}
\]

where \( i \) is the sequential number of clustering theme and \( k \) is the sequential number of clustering sample.

We will furnish a spatial clustering example composed of two themes \( T_1 \) and \( T_2 \). Suppose \( T_1 \) has three dimensions of \( f_1, f_2 \) and \( f_3 \), \( T_2 \) has two dimensions of \( f_1 \) and \( f_2 \). Suppose there are totally five clustering samples for convenience. For all samples, the weight value vector for cluster theme 1 is:

\[
Q_1 = \begin{bmatrix}
q_{11} \\
q_{21} \\
q_{31}
\end{bmatrix} = \begin{bmatrix}
0.31 \\
0.29 \\
0.40
\end{bmatrix}
\]

Let the dimension vector of Sample 1 is: \( W_1 = \{ f_{11}, f_{12}, f_{13} \} = \{ -0.31, -0.50, 0.81 \} \), where the value is acquired by investigation of domain expert. The actual value for each dimension of samples is normalized. Then we will get the calculated value of the Sample 1 for Theme 1; \( a_{11} = W_1 \times Q_1 = 0.0829 \). Similarly, the value \( a_{1k} \) of Sample \( k (k < 5) \) for Theme 1 can be calculated. By the same method, the calculated value of Sample 1 for Theme 2, \( a_{12} \), can be calculated and so is the rest samples. The calculation value matrix \( A_{12} \) can then be formed which includes calculated value elements for all samples and themes.

According to the principle of systematic cluster, to get clustering result for spatial samples, the basic principle is; firstly set all clustering samples (suppose the total number is \( m \) as one group), then divide the most dissimilar sample groups based on the similarity matrix of clustering theme. Similarly, another turn will be pro-
cessed and the most dissimilar group will be divided into new groups until all samples are divided into desired groups. It can be seen that the process of spatial clustering is dynamic and the group numbers are not set in advance.

2.5 Granularity adjustment and algorithm

How many groups should be divided for \( n \) number samples under a certain clustering space. Without special requirements, the rule of \( C_{\text{max}} \leq \sqrt{n} \) is reasonable, where \( C_{\text{max}} \) is the maximum number for clustered groups. This paper will follow this principle to reduce meaningless and time-consuming clustering for low granularity.

The number of clustering groups will vary between the integer number of 1 and \( C_{\text{max}} \). The clustered group number reflects clustering result on different level. Let clustering group number be \( c \) and \( 1 \leq c \leq C_{\text{max}} \), the procedures of result comparison for spatial clustering on attribute and geometric themes will be:

1) constructing geometric and attribute dimensions. Select meaningful clustering dimensions to form geometric and attribute, respectively. In the plane space, the geometric theme is coordinates \( x \) and \( y \), and in the solid space it is coordinates \( x, y \) and \( z \). The attribute theme is variable for different applications. The weight value for each dimension in each theme should be also defined.

2) normalizing every dimension value of clustering samples to erase the influence of error caused by different units of dimensions. A normalized geometric theme \( S(ID,x,y,z) \) and an attribute theme \( A(ID, \text{attribute 1}, \text{attribute 2}, \ldots, \text{attribute n} \) are formed.

3) calculating the possible maximum clustered groups \( C_{\text{max}} \).

4) recycling the clustered group from 1 to \( c \) (\( c = C_{\text{max}} \)). Suppose the current group number is \( t \), then:

- recursiving clustering to form clustered groups based on the similarity matrix until the clustered group number account to \( t \) for each theme, viz. possible clustered group number goes from 1 to \( t \),
- making distribution statistic of clustered samples for all \( t \) possible clustered results of the two themes separately and recording each sample number,
- analyzing the relative or inclusive index for clustered group numbered from 1 to \( t \) for the compared clustering theme,
- outputing the statistical result in form of table or graph,
- obtaining rules or knowledge according to the relative or inclusive index.

5) end.

3 Case study

3.1 Data preparation

Survey the soil distribution in an investigated region and digitalize the distribution with Arc’Info. Four layers in soil character \( f_1, f_2, f_3, \) and \( d \) are generated. The layer of \( f_1, f_2, f_3 \) represents the effective nitrogen of soil, effective phosphor of soil and soil texture, respectively. Then make AMI “UNION” operation on \( f_1, f_2, f_3, \) and \( d \) form multi dimension spatial samples (polygon entities), with each sample described by the attributes of \( f_1, f_2, f_3, \) and \( d \). Each sample is the integration of spatial location (reflecting the distribution of soil type), soil effective nitrogen, effective phosphor and soil texture. To firm the correlation of spatial distribution of soil types and soil attributes, the centre of clustering samples is extracted to represent the spatial location of each soil type, viz. the geometric theme. The other characters of soil (soil effective nitrogen, effective phosphor and soil texture) are the three dimensions that form the attribute theme. The coordinate \( x \) and coordinate \( y \) for geometric theme have equal weight 0.5. The dimensions for the attribute theme are also assigned an equal weight. All the dimension...
sions are normalized as experimental data.

3.2 Result comparison between clustering themes

The inclusive index is used here to measure the reliable degree of attribute character upon spatial location. According to the similarity evaluation index and clustering procedure discussed in Section 2.4 and Section 2.5, the result is shown in Fig. 3 and Table 1, where \( t \) is the recursive time, and \( v \) is the inclusive index of the attribute theme to geometric theme. From Fig. 3, we can see that \( v \), as a relation reflection of different themes, changes irregularly with the increase of clustering number \( t \). This indicates that there is no intrinsic relation between the inclusive index and the clustered number.

From Table 1, we can see that the themes used as clustering standards are geometric and attribute themes. The total samples clustered are 49,001. The index to evaluate the relation between the two themes is inclusive index and there is no obvious trend of inclusive index value with the increase of clustered number in Table 1.

| ID | Clustered number | Clustered result | Distribute of samples | Inclusive index \((v)\) \% |
|----|------------------|------------------|-----------------------|--------------------------|
| 1  | 1                | class 1          | 19,001                | 19,001                  | 100                      |
| 2  | 2                | class 1          | 29,144                | 26,273                  | 76.9                     |
| 3  | 2                | class 2          | 19,857                | 22,728                  | 42.1                     |
| 4  | 3                | class 1          | 16,936                | 11,365                  | 71.2                     |
| 5  | 3                | class 2          | 17,029                | 20,117                  | 55.1                     |
| 6  | 3                | class 3          | 15,036                | 14,319                  | 90.6                     |
| 7  | 4                | class 1          | 9,823                 | 6,890                   | 56.3                     |
| ...|                  |                  |                       |                         |                          |
| 116| 18               | class 2          | 8,439                 | 8,033                   | 80.2\%                   |

If a special point value 75\% is set as \( v \) standard for the inclusive index value, then the records can be filtered, as shown in Table 2. As it is always true that the inclusive index value is 100\% if the clustered number is 1, the first record in Table 2 is meaningless. The second record, however, does reflect that samples in Class-1 clustered based on attribute theme are 76.9\% included in those clustered based on the geometric theme when the total clustered number is 2. The same result is true when the total clustered number is 116. If the value of the special point is set lower, the more records will be retrieved.

| Clustered number | Clustered result | Distribute of samples | Inclusive index \((v)\) \% |
|------------------|------------------|-----------------------|--------------------------|
| 1                | class 1          | 19,001                | 19,001                  | 100                      |
| 2                | class 1          | 29,144                | 26,273                  | 76.9                     |
| 18               | class 2          | 8,439                 | 8,033                   | 80.2                     |

If the selected themes are correlative, it is possible to obtain association knowledge between different themes by analyzing the distribution character of clustered samples. In this case study, we get 29,144 samples on the geometric theme and 26,273 samples on the attribute theme when the total clustering number is 2. This shows that the spatial soil distribution partially
influences the soil character, and this rule can be used in soil classification and mapping research.

4 Conclusions

Knowledge discovered from spatial databases has been recognized as valuable knowledge acquisition in environment management, resource utilization and planning of industry and agriculture. This paper proposes the principle of result comparison for different clustering themes.

The general process of spatial clustering is completed when spatial entities are clustered into several groups with a unique theme as clustering standard. This paper continues this work by comparing those clustering groups done by different but related spatial clustering themes. Some definitions and a set of calculating indexes, viz. relative index and inclusive index, are defined to facilitate the research, and then an algorithm to complete the spatial clustering result comparison between different clustering themes is presented.

Our research shows that some valuable spatial correlation patterns can be further found from the clustering result comparison with multi-themes, based on traditional spatial clustering as the first step. Those patterns can tell us what relations those themes have, and thus will help us have a deeper understanding of the researched spatial entities and it is proved by the case study.

REFERENCES

1. Leung Y., Ma J. H., Zhang W. X. (2001) New method for mining regression classes in large data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:5-21
2. Lu W., Han J., Oci B. C. (1993) Discovery of general knowledge in large spatial databases. The Far East Workshop on Geographic Information Systems, Singapore.
3. Taylor K., Walker G., Abel D. (1999) A framework for model integration in spatial decision support systems. *Int. J. Geographical Information Science*, 13: 533-555
4. Aldenderfer M. S., Bashfield R. K. (1984) Cluster analysis. Beverley Hills: Sage Publications.
5. Pugh T. W. (2003) A cluster and spatial analysis of ceremonial architecture at Late Postclassic Mayapan. *Journal of Archaeological Science*, 30: 941-953