Chapter 2

Research on Spatial Data Mining in E-Government Information System

Bin Li, Lihong Shi, Jiping Liu and Liang Wang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/50245

1. Introduction

E-Government Information System is the idiographic application of GIS in the field of government departments in the world[1]. There are large amounts of data stored in the database of E-Government Information System. 80 percent of the data is concerned with spatial location. In fact, there are little applications of these data. A great deal of the data is idle, which has caused a huge waste of data due to rarely effectively utilization in practice. Actually, Spatial Data Mining, i.e.SDM, is a kind of important and useful tool in the practical application of E-Government Information System database, and is very useful to find and describe a hidden mode in the particular multi-dimensional data aggregation. It is very necessary to deal with the task of spatial data mining based on E-Government Information System database with different data resources, data types, data formats, data scales. "Mass spatial data and poor knowledge" has become a gap for the development of geo-spatial information science, which requires necessary data mining. SDM can mine automatically or semi-automatically unknown, creditable, effective, integrative or schematic knowledge which can be understood from the increasingly complex spatial database and enhance the ability of interpreting data to generate useful knowledge. And with time goes on, from its beginning, SDM has attracted more and more attention, and achieved some academic and applied results in the field of artificial intelligence, environmental protection, spatial decision-making support, computer-aided design, knowledge-driven image interpretation, intelligent geographic information systems (GIS), and so on.

Although SDM is a kind of important tool in the practical application, it is very difficult to find information manually because the data in the data set grows rapidly. The algorithms on data mining allow automatic mode search and interactive analysis.

This chapter consists of six components. The first section is the introduction. Section 2 provides the data characteristics analysis of E-Government Information System. The third section

describes the course of spatial data mining. And section 4 presents the examples of spatial data mining. The next section gives the conclusions. And the last one is the acknowledgement. Especially, in this chapter, a useful application example on land utilization and land classification in a certain county of Guizhou Province, China, is presented to describe the course of SDM. Thereinto, a derived star-type model was used to organize the raster data to form multi dimension data set. Under these conditions, clustering method was utilized to carry out data mining aiming at the raster data. Based on corresponding analyses mentioned in this chapter, users could find out which types of vegetation were suitable for being cultivated in this region by using related knowledge in a macroscopic view. Therefore, feasible service information could be provided to promote economic development in the region.

2. Data characteristics analysis of E-Government Information System

The construction of E-Government Information System is complex with the character of being Systemic, which involves different fields. And the data used in the system is also complex with many types. In a whole, the data character can be described as follows.

1. Diversity of data with multi sources
   The data used in the E-Government Information System mainly comes from different departments at different level. And it includes various kinds of basic geographical data and monographic data with different scales. According to the spatial partition of data type, it includes DLG, DRG, DEM and remote sensing image with multi spectrum, scale, time, etc. And non-spatial data has the different types of statistical information, multimedia text, image, video, and sound, etc. The multi types or forms of data format, data represent, database structure and data store have been produced owe to the diversion of data source.

2. Great amount of data content
   There are large amount of different kinds of data stored in the database of E-Government Information System. Thereinto, the majority part of data capacity is spatial data, especially the image, DEM data. Besides, the video content belonging to non-spatial data is also very large. With the development of new information revolution, E-Government will face new demands. And more and more new data will be produced to meet the need or require of E-Government Information System. Meanwhile, those old data must also be restored in the database as the history data in case of recovering.

3. Temporal character of data change
   With the development of E-Government Information System, the data used in the system is constantly extended with more and more amount, content and type. Therefore, the relational data is often changed and update with temporal character. Temporal sequence is also the main character of spatial data, which is often used to make the comparison to show the development of the task.

4. Spatiality character
   Spatiality is the main character of spatial data compared to non-spatial data, which includes mainly geographical location and form of spatial object. The former often

includes geographical coordinate, postal code, administrator code, toponym, address, etc.

3. Process of spatial data mining

The detailed process of SDM can be divided into five phases described as follows: demand investigation, data selection, data pre-processing, data conversion, data mining, knowledge representation and evaluation. In figure 1, the flow of SDM can be described.

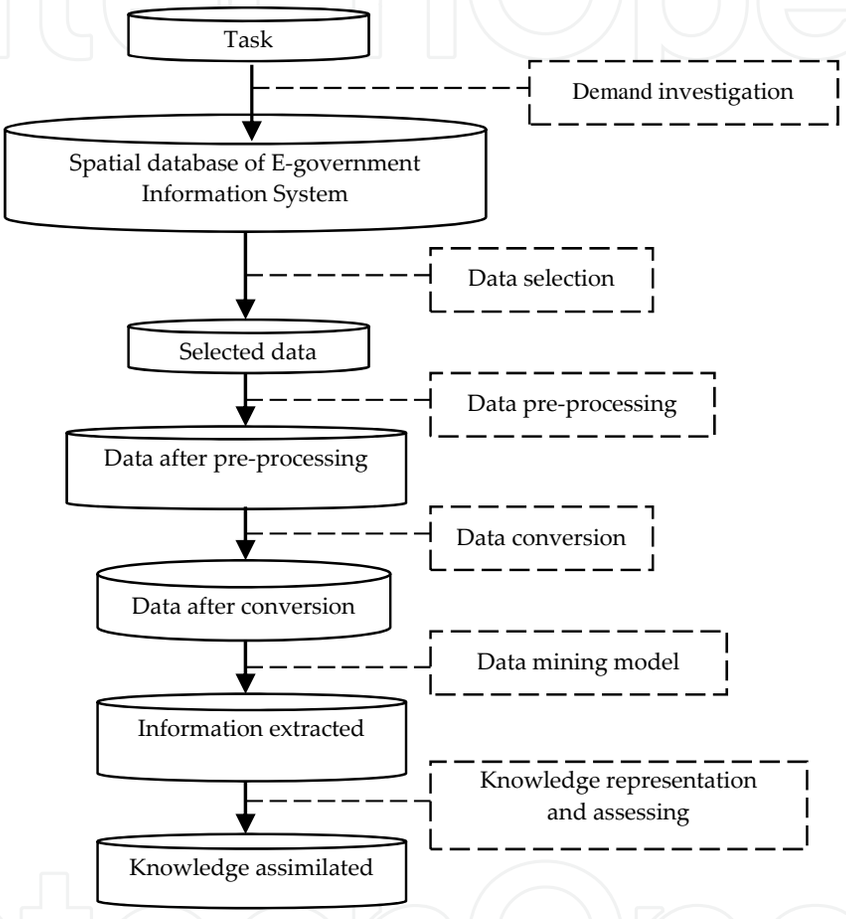

Figure 1. Process of SDM

3.1. Demand investigation

It is necessary to understand the existing data and business information before data mining. Understand fully the problems to be solved and give a clear definition on the goal of data mining. Therefore, demand investigation is the necessary and first step of SDM according to the task oriented to actual application.

3.2. Data selection

Data selection is carried out after finishing demand investigation and knowing unambiguous demand. Data selection is the course of ascertaining data source to be used in

data mining and collecting data records stored in the database in accordance with the standards of being established. Some rules or methods of data selection must be made or adopted to select the needed data during the process.

3.3. Data processing and conversion

After the completion of data selection, it is necessary to make data pretreatment. The Main work in this step is to carry out data cleaning aiming at the data stored in the database of E-Government Information System and delete unnecessary information[9], and transform the required data into a unified data format. Besides, data conversion will be made through the process of data merger and integration to convert them into data with identification after data pretreatment.

3.4. Data mining model construction

It is the important step to construct the data mining model. Corresponding data mining models are to be constructed aiming at different demands, such as decision tree, clustering, etc. It is decisive for data pretreatment to select a certain type of model. Data mining can be made aiming at the data stored in the database of E-Government Information System and pattern extraction can also be done after determining the data mining model.

3.5. Knowledge representation and assessing

The results should be interpreted when data mining is completed. During the course, some of the process may be returned to the front processing steps in order to obtain more efficient knowledge. And knowledge extraction can be made repeatedly so as to gain more effective information to be interpreted as the knowledge for decision-making in the future. Finally, performance of the applied model needs assessing and constantly improving the algorithm to meet the needs of different actual application[13].

4. Realization of spatial data mining

There are lots of microcosmic data existing in the E-Government Information System coming from different sectors and different periods of time, which mainly includes vector data, raster data, attribute data and a lot of statistical data, etc[2]. It is necessary for the above-mentioned data to be stored in the multi dimension data set in accordance with multi topics to achieve the course of fixed and random dynamic data query processing, comprehensive analysis. Users can access the multi dimension data set in the end of client and extract correlative data from the data set in the background according to different rights, and selected theme demands and finally generate a variety of visual results, such as various statistical graphics (histograms, pie charts, pyramid diagram, etc.) and statistical analysis of statements, and so on. Meanwhile, some corresponding multi-dimensional analysis aiming at disposal results, such as cluster analysis, can also be made. In the next section, the raster data mining will be made as an example to prescribe the whole process[10,11].

4.1. Raster data processing and importing

Raster data roots in 1:250,000 DEM data cropped in a certain region of Guizhou Province, China. Grid processing is completed by using grids (each grid represents 100 meters x 100 meters) in the region. A pivot can be got in each grid and imported into the relational database to be a record. The picked data includes grid number, terrain slope, terrain aspect, grid coordinates, land use type and land cover type, etc. There are more than 280 records generated in the region.

Slope and aspect, which are correlative each other, are parameters commonly used to describe the terrain. Slope reflects the inclination degree of slope and the latter represents the direction which the slope faces. In terms of land use and land cover, the slope and aspect have many uses. For example, the slope exceeding 35 degrees should not be developed in general in the agricultural land development. Slope is the function of points, which is the angle between the normal direction N and vertical direction Z aiming at a certain point in the surface and can be represented using α. In practice, it is no use in computing the slope of each point[2]. Average slope of a basic grid unit is often used to represent corresponding signification. Aspect is the angle between the projection of normal direction and the due direction, that is, the direction of azimuth between the projection vectors. B is often to be described to represent the aspect. See fig.2.

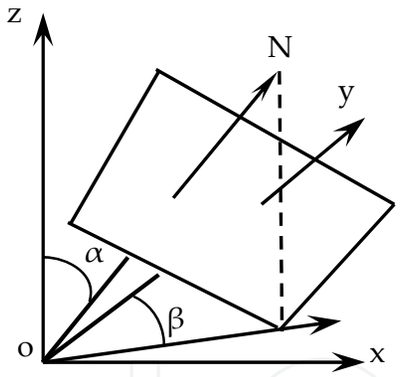

Figure 2. Terrain slope and aspect

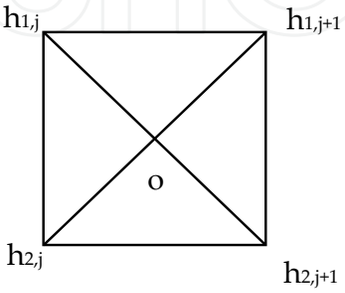

Figure 3. Elevation representation

$$\alpha = arctg\left(u^2 + v^2\right)^{1/2}, u = \left(h_2 - h_3\right) / \left(2^{1/2} * d\right), v = \left(h_1 - h_4\right) / (2^{1/2} * d)$$

Thereinto: *h1*, *h2*, *h3*, *h4* is respectively the elevation of four corner points in the grid, *d* is the grid length of each side.

$$\beta = tan^{-1}\left(dx * a_j / dy * b_j\right), a_j = h_{1,j+1} + h_{2,j+1} - h_{1,j} - h_{2,j}, bj = h_{2,j} + h_{2,j+1} - h_{1,j} - h_{1,j+1}$$

Similarly, *dx*, *dy* represents respectively grid length of horizontal or vertical side, $h_{1,j+1}$, $h_{2,j+1}$, $h_{1,j}$, $h_{2,j}$ denotes respectively the elevation of four corner points in the grid. See fig.3.

The created content of data during the course of being cropped can be described as follows, which takes slope for example. See fig.4.

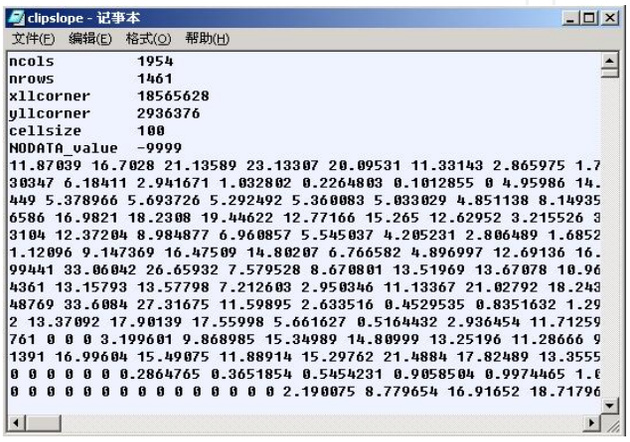

Figure 4. Slope data in a certain region

In Fig.4, *ncols* represents the lists of the raster and *nrows* shows the rows. And *x11corner*, *y11corner* represent respectively the coordinates of left underside corner in the whole raster range. Besides, *cellsize* denotes the grid size, which can be expressed in the manner of meter. Corresponding values can be got by using procedures to import the above-mentioned data into the relational databases of SQL Server2000. These values may be arranged in the sequence of known row or list number. For example, corresponding value is *(i, j)*, *(i, j +1)* respectively when *id* equals 1, 2. Similar conclusions can also be made. The arrangement method of raster data can be described in table 1 and figure 5.

i,j	i,j+1	i,j+2	i,j+3	i,j+4	……	i, ncols
i+1,j	i+1, j+1	i+1, j+2	i+1, j+3	i+1, j+4	……	i+1, ncols
i+2,j	i+2, j+1	i+2, j+2	i+2, j+3	i+2, j+4	……	i+2, ncols
i+3,j	i+3, j+1	i+3, j+2	i+3, j+3	i+3, j+4	……	i+3, ncols
……	……	……	……	……	……	……
nrows, j	nrows, j+1	nrows, j+2	nrows, j+3	nrows, j+4	……	nrows, ncols

Table 1. Arrangement method of raster data

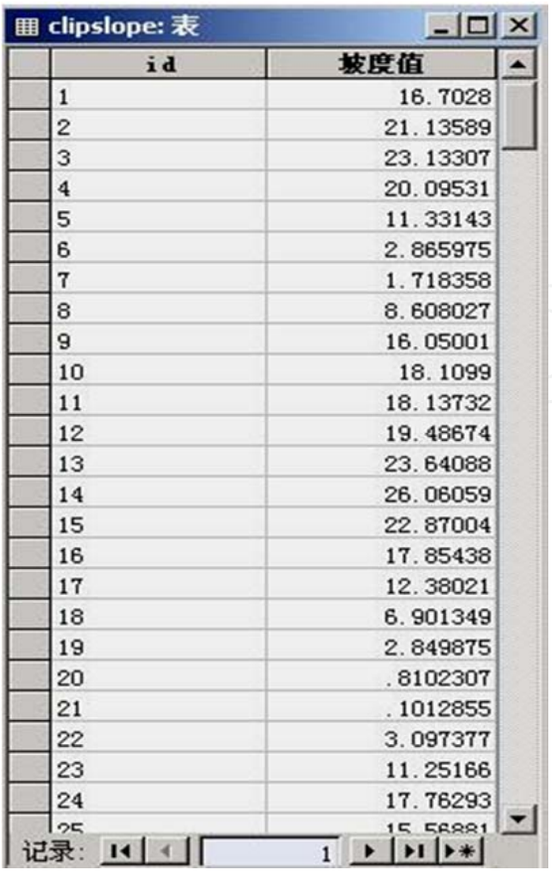

Figure 5. Arrangement method of slope

4.2. Construction of derived star-type model based on raster data

Each star-type model includes a fact table and some corresponding dimension tables. All the fact tables and dimension tables are stored in the SQL Server2000 database. The key of constructing a star-type model is to design right fact table and dimension table as well as establish mutual connections between them[14]. These tables can reflect the complex relationship among different data. Furthermore, too much redundant data can be produced if using one dimension table to describe due to corresponding complexity of relationship and data in the dimension in practice. Here, the dimension can be divided once again. Some branches will appear from the angle of "star", thus, a derived star-type model, which is also called snowflake-type model, would come into being [2]. The derived star-type model established in this section has a specific theme on land use and land cover. Here is an example happened in a certain region of Guizhou Province.

4.2.1. Construction of fact table

In the fact table, i.e. Combine Table, it mainly stores the main code, which is represented with gridid, and the values of slope and aspect in each grid. The latter can be extracted automatically from DEM by making use of relevant designing program. Designed fact table can be described as follows in figure 6.

Figure 6. Designed fact table

4.2.2. Construction of dimension table

Dimensional tables mainly include three tables, thereinto, dimensional table of grid can store each grid's main code used to link fact table, row number, list number, coordinates and the codes of land use and land cover. The dimensional table of land cover can store corresponding attributed value, such as land cover denomination, region name. The dimensional of land utilization, similar to land cover table, stores the attributed value corresponding with grid table, such as land utilization denomination, region name, etc.

Dimensional tables of *land cover*, *land utilization*, *grid* can be described respectively as follows in fig.7, fig.8 and fig.9.

Figure 7. Dimension of land cover

Figure 8. Dimension of land utilization

4.3. Construction of multi dimension data set based on raster data

Compared with traditional approach, the main difference of constructing multi dimensions data set based on raster data in this section is the dissimilarity of forming star-type model.

Necessary data grouping processing can be adopted in order to improve the system function because the raster data amount is quite larger. In this section, all the raster data can be divided into more than 700 groups. Data of each group may arrange in the sequence of keyword.

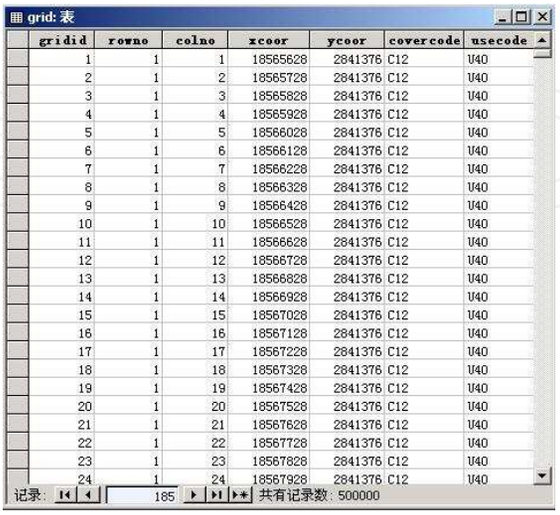

Figure 9. Dimension of grid

New multi dimensions data set can be constructed after adopting snowflake-type framework. As shown in figure 10, dimensions of *land cover* and *land utilization* act as the branches of dimension of *grid*.

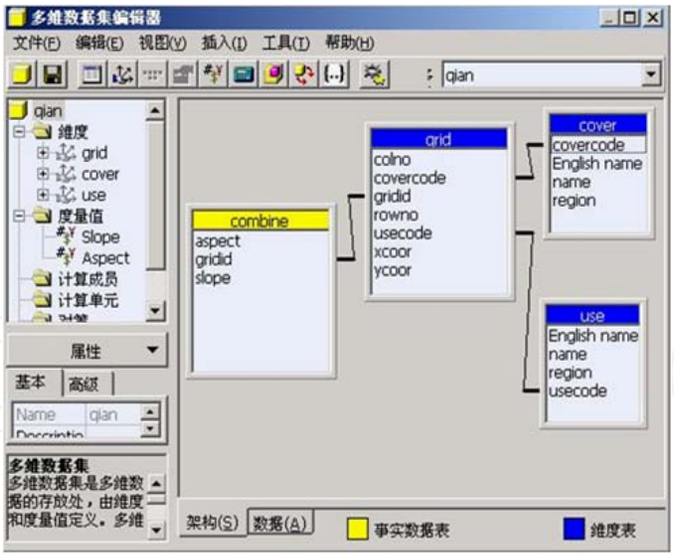

Figure 10. Multi dimension data set

4.4. Data mining based on raster data

Process of data mining based on raster, similar to the one of data mining based on vector, has also included model creation, example dimension choice and forecasting the entities and data to be trained as well as the model processing, and so on.

4.4.1. Outline of spatial data mining algorithm

At present, there are many spatial data mining algorithms, such as statistical analysis, neural networks, clustering, decision trees, genetic algorithm, classification, etc. In this section, clustering algorithm is mainly used to deal with the raster data stored in E-Government Information System. In the processing course of clustering algorithm, diverse data can be divided into different categories in order to make the difference between categories as big as possible and inner differences in the category as small as possible[4]. Clustering algorithm, which is also called aggregation algorithm, is an indirect data mining algorithms and does not use independent variables to get designated output. Different from classification model, clustering algorithm does not know beforehand that there are several categories to be divided and what these categories are. And it does not know how to define these categories according to some data items. Although it cannot predict unknown data value, while classification algorithm can, it provides a way to find similar records. These records can be considered the components of given clusters according to self-determined algorithm itself[3]. The following section has described the course of data mining by using constructing clustering.

In this section, data mining is made based on the algorithm of EM. The algorithm is one of spatial clustering algorithms, which adopts the probability method of distribute each data to the determined classification instead of calculating directly the distance.

Set up a closed curve in each dimension as the clustering criteria, and calculate the mean and standard variance. If the point falls within the closed curve, then they have a certain probability of belonging to the given classification. Since the different curve of corresponding classification is not only, so the data point may also fall in multiple classification curves and is given a different probability[6,7]. During the course of implementation, the initial classification model is repeatedly made to finish the step of optimization by the algorithm to fit the data, and determine the probability of data point existing in a classification. When the probability model fits the data, the algorithm terminates the process[19]. The function to determine the suitability is the logarithmic likelihood data fitting the model. In this process, if an empty classification is generated or the membership of one or more categories is less than a given threshold, new data point will be reseed in the classification with low fill rate, and the algorithm will re-run.

As the results of the EM cluster analysis algorithm is probabilistic, which means that each data point belongs to all categories[20], and calculates a probability for each combination of data point and the classification, but the data point is allocated to the classification of each with a different probability.

This method allows the classification overlap, so the total number of items in the all categories may exceed the total number of items in the training set[8,12]. Aiming at the mining model results, the instructions to support the scores can be adjusted accordingly in order to illustrate this situation.

4.4.2. Merit of the algorithm

Compared to the traditional clustering algorithms, the algorithm has several merits.

1. The algorithm is better than the sampling algorithm with foretype.
2. The algorithm has better efficiency in working, which scans the database only once.
3. There is little influence in executing the algorithm.

4.4.3. Implementation of the algorithm

In brief, the algorithm includes two steps as follows.

Step 1. Estimation.
Step 2. Maximization.

Suppose existing n sample (n>0), which come from K Gaussian Mixture Distribution. Each distribution is mutually independent[21].

The parameter of Gaussian Mixture Distribution can be estimated to determine the Mean Value, Variance, prior probability of each distribution.

The probability density function can be described as follows.

$$n_j(\vec{x},\vec{\mu}_j,\sum_j)=\frac{1}{\sqrt{(2\pi)^m\left|\sum_j\right|}}\exp[-\frac{1}{2}(\vec{x}-\vec{\mu})^T\sum_j^{-1}(\vec{x}-\vec{\mu})]$$

Thereinto, $\vec{x}$ is observed data vector, $\vec{\mu}$ is the corresponding data mean. And i, j is the different data object, m is the variance.

The priori probability meets:

$$\sum_j \pi_j = 1$$

The maximum likelihood estimation is expressed as follows.

$$P(\vec{x}_i)=\sum_{j=1}^{k}\pi_j n(\vec{x}_i;\vec{\mu}_j,\sum_j)$$

Suppose $\theta_j=(\vec{u}_j,\sum_j,\pi_j)$ is the parameter of the jth Gaussian Mixture Distribution, and the parameter space to be estimated can be described as follows.

$$S=(\theta_1,\cdots,\theta_k)^T$$

The probability formula of sample X is

$$l(X\left|\ S\right)=\log\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j n_j(\vec{x}_i,\vec{\mu}_j,\sum_j)$$
$$=\sum_{i=1}^{n}\log\sum_{j=1}^{k}\pi_j n_j(\vec{x}_i;\vec{\mu}_j,\sum_j)$$

Continuously make the steps of iterative estimation and maximization, and compute repeatedly the above-mentioned three parameters until $l(X \mid S)$ meets the condition of no significant increase again[22].

4.5. Example of spatial data mining

An idiographic example on land utilization and land cover, which happened in a certain region of Guizhou Province, can be used to describe how to create a mining model. Land utilization and land cover are the most prominent landscape signs in the earth surface system. Land utilization is a process of making natural state of the land into the one of artificial ecosystems. And the latter represents objective existence of the earth's surface with specific spatial and temporal attributes. Generally, land types can be identified indirectly according to combination of land cover and structure. The created data mining model can be described as follows in figure 11.

Figure 11. Data mining based on clustering

As shown in figure 11, the original data set is divided into three categories. Thereinto, *C6*, *C8*, *C12* represent respectively land cover type of lush shrubbery, flourish grassland and cropland. The three land cover types account for the vast majority of the total number of all types. The amount is about 96.15%. Simultaneously, other land cover types had little proportion and can be almost negligible.

In-depth analysis to *Cluster 1* node can be made. As shown in figure 12, in this category, most of land cover types are *C6* and *C8*. The proportion is about 95.51%. That is, the type of land cover in this region is basically shrubbery and grassland. It is more suitable for the development of animal husbandry in a view of economic factor.

Click the node feature set *usecode*, the data in the table can be changed accordingly. As shown in figure 13, all the land utilization type appeared in the *Cluster 1* node is *U10*, which represents the forest. By adopting similar steps, corresponding results can be shown in the whole region included by *Cluster 1* node. For example, the average slope and aspect are respectively 12.51, 184.22 degree. Based on these analyses, users can find out which types of

vegetation are suitable for being cultivated in this region by using related knowledge in a macroscopic view[16,17]. Besides, light conditions may also be considered to improve the development of vegetation. Therefore, feasible service information can be provided to promote economic development in the region.

Figure 12. The dominant type of land utilization

Figure 13. The single type of land utilization

Similarly, analysis to node *Cluster 2* or *Cluster 3* is analogous to the node *Cluster 1*.

In another view, distribution information in the whole categories can be found in accordance with node feature set characteristics on the contrary. Likewise, the distribution of other node feature sets characteristics is similar to the above-mentioned instance. As shown in figure 14, it represents the type of land cover of *C6*.

In figure 14, the most part of land cover type *C6* belongs to the code *Cluster 1*. And if the land cover type is switched from *C6* to *C12*, the most part of land cover type *C12* will belong to the code *Cluster 2*, the proportion of which is about 97.7%.

The distribution character of other node feature set is similar to the above-mentioned part.

Figure 14. The type of land cover of *C6*

5. Conclusions

In this chapter, spatial data and corresponding attribute data of E-Government Information System were mined deeply by applying the technology of data mining such as Cluster. And some useful information and knowledge were extracted. For example, the correlative relationship of land utilization and land classification could be found. It was convenient for the departments of different level to make aided decision-making. Although the different applications of SDM have been increasingly extended than ever, there are still some limitations on applicability and efficiency existing in the actual application system[5,15]. At the same time, SDM is also a new study field with more attraction and challenges[18]. With the enhancement of information and the development of soft, hardware and other technology, SDM can possess mighty knowledge discovery function, and can still effectively bring into play known or potential value especially in E-Government Information System.

Author details

Bin Li, Lihong Shi, Jiping Liu and Liang Wang
Chinese Academy of Surveying and Mapping, Beijing, China

Acknowledgement

The work presented in this chapter has been funded by Fundamental Scientific Research Foundation (No.7771204, No.G71012, No.77738). And all the data used in this chapter comes from the database of Government GIS. Thank all my colleagues for their enthusiastic supports and helps.

6. References

[1] Qingpu Zhang, Foxiao Chen(1999) Basic conception and construction mode of Government GIS. Remote Sensing Information, supplement,

[2] Bin Li(2002) Research on construction of Governmental GIS Spatial Data Warehouse. The master's paper of Bin Li's published by Chinese Academy of Surveying and Mapping.
[3] Bin Li(2002) Study of Construction and Application of Data Warehouse for Government GIS. Bulletin of Surveying and Mapping, 2002(2): 4-6
[4] Bin Li, Lihong Shi, Jiping Liu(2009) A method of raster data mining based on multi dimension data set. The 6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD
[5] ChinaKDD(2008) Schedule of clustering arithmetics on data mining, http://www.dmresearch.net
[6] Li Deren, Shuliang Wang, Wenzhong Shi, Jizhou Wang(2001) On Spatial Data Mining and Knowledge Discovery. Wuhan University Transaction, Vol 26(2):491-498
[7] Genlin Gi, Zhihui Sun(2001) Jou rnal of Image and Graphics. Vo l. 6 (A) , No. 8: 715-721
[8] Xingli Li, Peiun Du,etc(2006)SClFNCE & TECHN0LOGY lnformation, Vol (6) : 158-160
[9] Miller H J, Han J(2001) Geographic Data Mining and Knowledge Discovery. London and New York:Taylor and Francis
[10] Wang S L,Wang X Z,Shi W Z(2003)Spatial Data Cleaning. In:Zhang S C,Yang Q,Zhang C Q, Terrsa M. Proceedings of the First International Workshop on Data cleaning and Preprocessing. Maebashi City, Japan:88-98
[11] Deren Li(2002) Spatial Data Mining and Knowledge Discovery Theory and Methods. Journal of Wuhan Univsrsity-Information Science, Vol(27) : 221-233
[12] Tongming Liu9(2001) Technology and application of data mining. Beijing: National Defence Industry Press
[13] Shengwu Hu(2006) Quality assessment and reliability analysis. Beijing: Surveying and Mapping Press
[14] Hernandez M A, Stolfo S J(1998)Real-world data is dirty: data cleaning and the merge/purge problem. Data Mining and Knowledge Discovery,2: 1-31
[15] Aspinall R.J,Miller D R., Richman A R(1993)Data quality and error analysis in GIS: measurement and use of metadata describing uncertainty in spatial data. In Proceedings of XIIIth Annual ESRI User.Palm Springs: 279-290
[16] Pawlak Z (1997)Rough Sets.In:Lin Y,cercone N.Rough Sets and Data Mining Analysis for Imprecise Data. London: Kluwer Academic Publishers: 3-7
[17] Kaichang Di(2000) Spatial data mining and knowledge discovery. Wuhan: Wuhan University Press
[18] Kelleller A, Abbaspour K, Schulin R(2000) Uncertainty Assessment in Modelling Cadmium and Zinc accumulation in Agricultural Soils. In:Heuvelink G B M, Lemmens M J P M. Accuracy 200:Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resource and Environmental Science. Amsterdam,The Netherlands:University of Amsterdam: 347-354
[19] Sheikholeslami G, Chatterjee S, Zhang A(1998) Wave-Cluster: A multi-resolution clustering approach for very large spatiall databases. In: Proceedings of the 24th International Conference on Very Large Databases. New York, 428-439
[20] Caiping Hu(2007) Spatial data mining research review. Computer Science, vol (5): 14-17

[21] Clementin E,et al (2000) Mining Multiple Level Spatial Association Rules for Objects with a Broad Boundary. Data and Knowledge Engineering, vol(34): 251-270
[22] Kacar E,et al(2002)Discovery Fuzzy Spatial Association Rules,Data Mining and Knowledge Discovery: Theory Tools and Technology IV. In: Dasarathy B V, ed. Proceedings of SPIE, Vol(4730): 94-102

IntechOpen