Design and Implementation of Spatial Data Mining System (M-SDM) based on MATLAB

Zhao Lu
China University of Geosciences, Beijing, China
Email: zhaolu1985cn@yahoo.com.cn

Zheng Xinqi
China University of Geosciences, Beijing, China
Key Laboratory of Land Regulation, Ministry of Land and Resources, Beijing, China
Email: zxqsd@126.com

Wang Shuqing
China University of Geosciences, Beijing, China
Email: w_sq_2002@163.com

***Abstract*—Taking the design of data analyzing software for imaging brain function — SPM (Statistical Parameters Mapping) for reference, this study combined MATLAB, GIS and SDM organically, constructed an SDM system framework on MATLAB platform, integrated the major algorithms such as spatial association rule mining, spatial clustering analyzing, decision tree analyzing and so on, and applied the system in land-use spatial database, aiming at enhancing the efficiency of massive data processing and enlarging the application of MATLAB in spatial data mining, spatial vector data processing and other aspects.**

***Index Terms*—spatial data mining system; spatial association rule mining; spatial clustering; decision tree; GUI; visualization**

I. INTRODUCTION

Through out the 1990s, the concept of Spatial Data Mining (SDM) was put forward. Since then, the data mining and knowledge discovery based on mass data have been paid more attention. There is mainly theoretical and pint-sized research among the current productions [1~8]. Nowadays, the research on data mining integrated system is few. However, this is a practical question to be solved.

MATLAB is famous for its outstanding data calculation and visual graphic representation function, which is superior to other software [9]. SPM is one of its successful examples for its advantage in GUI and mass data processing [10]. Its success proves the feasibility of developing data process and analysis software based on MATLAB.

This study took full advantage of the current database and visualization technology, constructed multi-arithmetic spatial data mining system based on MATLAB7.1, expecting to make helpful discussion on the disposing of mass data in SDM and the application of MATLAB in SDM, visualization and system construction.

This paper is sponsored by the national natural science foundation of China (No.40571119), the national social science foundation of China (No.07BZZ015) and the national key technology R&D program of China (No.2006BAB15803).

II. DESIGN OF M-SDM

M-SDM is the integrated SDM system, which is based on MATLAB and coded in M-language. It faces mass data and integrates multi-algorithm of SDM, and has excellent interaction with users.

A. Study purpose and demand analysis

On the basis of MATLAB and SDM fundamental theory, we take the practical needs of spatial information management and processing, SDM construction and result analysis into consideration, aiming to discuss the application of MATLAB in SDM and spatial vector data process, and to provide effective technological means for mass data process and visually evaluation in SDM.

Through kinds of communication with users, we found out the user demand is as following:

(1) The system can implement the mainly algorithms for SDM. Faced to mass data, users can complete the whole process of SDM by simple operation in friendly interface.

(2) During the SDM process, enhance the interactivity between users and the system. And users can gain the inter-results visually.

(3) The system should provide multiple exporting modes to exhibit the results in intuitive manner which is convenient for analyzing and decision-making.

B. Design of the architecture

M-SDM has a triplex architecture—data source, spatial data miner and user interface (see figure 1). Users select

© 2008 ACADEMY PUBLISHER

the related data through spatial database management tool directly, analyze the query results visually, and then extract the data which is related to the target domain. Afterwards, users analyze the extracted data by kinds of data mining methods to choose the appropriate methods. And then, users set the corresponding parameters interactively to mine the latent rules and patterns. The gained information and knowledge is fed back to users in a visual manner, so they can analyze and evaluate the knowledge in order to support spatial decision-making.

During the whole process of data mining, each tier is interpenetrated with others. Users can control each step. Moreover, we can get the final satisfying results by iterative man-machine interaction.

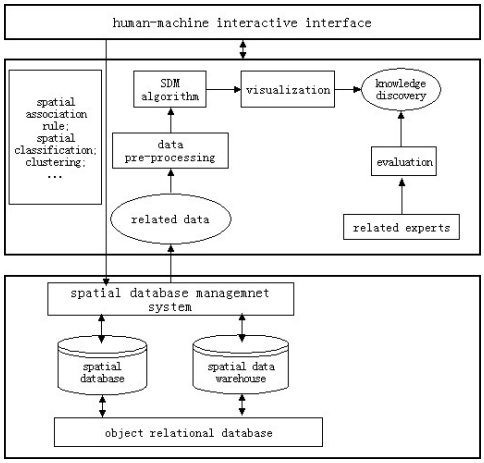

Figure 1. The architecture of M-SDM

C. Design of the main function

Based on the user demand and the design of system architecture, this system should have the following basic functions: 1) Visualization of the original data. The original data should be displayed in series of visualization technology and means, so the users can understand the distribution and statistics information of the spatial data well. 2) Data reduction. Carry out data processing and character extracting and choose partial data to represent the whole data set by wiping off the redundant or meaningless data. 3) Data pre-processing. Convert the original data into the target format of SDM algorithm, including data standardization, discretization and so on. 4) Knowledge discovery and visualization. We mine the latent knowledge and patterns by effective SDM algorithms, and feed them back to users by visualization technology.

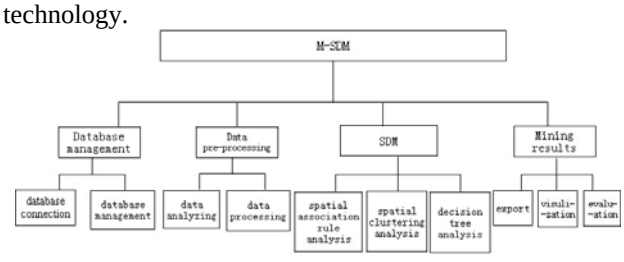

Figure 2. The functional structure of M-SDM

In general, the main functions of M-SDM include database management, data pre-processing, SDM and visual assessment. The functional structure of the system is shown in figure 2. In this system, we can carry out spatial association rule mining, spatial clustering analyzing and decision tree analyzing. Moreover, the three parts of SDM adopt Apriori, fuzzy clustering and C4.5 algorithm, respectively.

III. Design and Implementation of GUI

Under the support of Graphical User Interface Development Environment (GUIDE) in MATLAB7.1, according to the main interface character of current software, M-SDM offers several interactive operation such as form input, menu and direct interaction. The built-in dialog box can not only help to simplify the system development, but also make the interface more humanistic and interactive. The main involved ones include question dialog box, message dialog box, input dialog box and list dialog box, which are realized by questdlg, msgbox, inputdlg and listdlg function, respectively.

The main interface of M-SDM is comprised of three-module buttons and other assistant function buttons (see figure 3). Each module integrates database management, data pre-processing, specific SDM algorithms, visual expression and assessment, and other function. The interfaces of modules mainly include menu, toolbar, buttons, input section, assistant view section and information output section (see figure 5, 6 and 10).

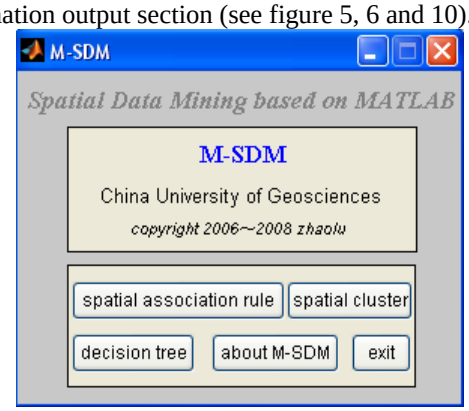

Figure 3. Main interface of M-SDM

IV. Implementation and Application of Main Function

A. Connection with database

In order to make full use of current database technology to manage and process data, we consider the characteristic of database connection in MATLAB, and adopt ODBC to realize the connection with database in M-SDM which is carried out mainly by calling Database Toolbox and Visual Query Builder. Figure 4 shows the interface of Visual Query Builder.

Besides, Database Toolbox of MATLAB offers some related functions to help connect and operate the database, such as connection, cursor, fetch, insert, update, set and so on. These functions can be used to supplement Visual

© 2008 ACADEMY PUBLISHER

Query Builder and perfect the interaction between M-SDM and database.

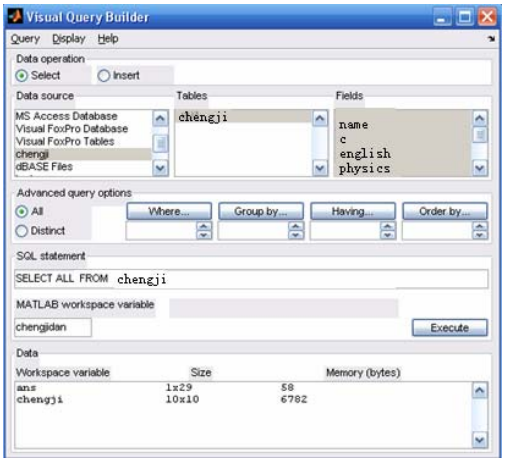

Figure 4. The interface of Visual Query Builder

B. Spatial association rule mining

Because of the technical methods and other limitations, people's knowledge of land use mode in different topographic forms is fuzzy. Nowadays, we pursue a high efficient utilization mode of land resource, so the fuzzy cognition must be improved under the support of advanced concept and technical method, then the comprehensive production ability of land resource may be exploited better. Under the support of spatial association rule mining module in M-SDM, we study and discuss the association rule between land use type and the corresponding land slope.

The study area is Ancheng village of Pingyin county in Shandong province in China. According to the China land classification criterion [11], there are 24 land use types in Ancheng which distribute in different topographic forms. The data comes from the production of Ancheng 1:10000 land use renewed survey database in 2004 and the 1:10000 geographical map of Ancheng in 2004.

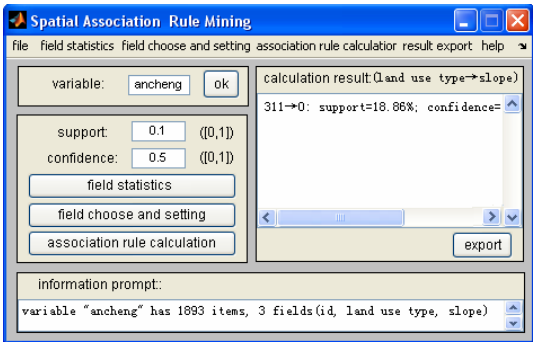

Figure 5. Interface of spatial association rule mining module

After being set the support threshold and confidence threshold as 0.1 and 0.5 respectively, the system calculates association rule----"land use type → slope" automatically. The gained rule is as following: in Ancheng, when land use type is grass ground and the land slope is below 2° , the support and confidence of this rule are 18.86% and 97.01%, respectively (see figure 5). According to our study [12], we find out that the area of land whose slope in Ancheng is below 2° accounts for 49.43% of the total area of this village while the area of grass land accounts for 20.14%, and the area of grass land whose slope is below 2° accounts for 88.89% of the total area of grass land. Thus it can be seen that there are some irrationalities of the land use in this village. The local land managers and decision-makers should make right land planning to guide local land use.

C. Spatial clustering analysis

Nowadays urban land grading is carried out in national and provincial scale [13, 14], but the land grading in country scale is weak and is of great significance to strength land management and perfect land grading [15]. We carry out the land grading of Changqing district in Jinan based on spatial clustering module in M-SDM, aiming to find out the regional differences of land quality and provide reference for land grading in country scale.

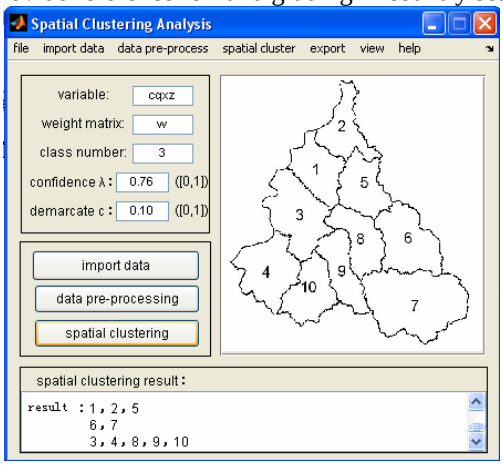

Figure 6. The land grading result

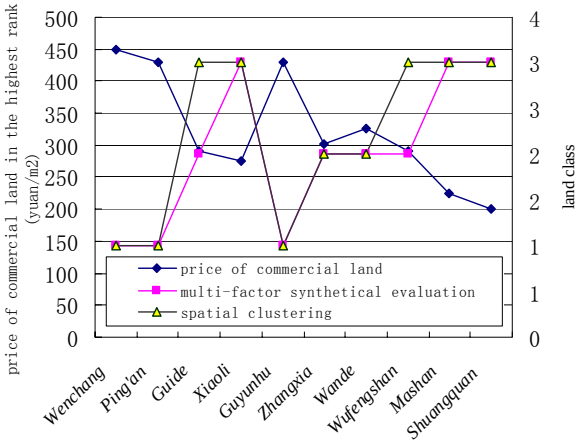

Figure 7. Comparison of the land grading results

Changqing is located in the middle of Shandong province in China, and it is 22km far away from Jinan city. It was authorized by State Department to become a district of Jinan but not a county yet. Now there are 10 offices, towns or villages under its jurisdiction—Wenchang, Ping'an, Guyunhu, Wufengshan, Guide, Xiaoli, Zhangxia, Wande, Mashan, Shuangquan. Based on the current production of land classification and

© 2008 ACADEMY PUBLISHER

grading and the acquirement of related data or information, we establish the land grading index system for Changqing district (see reference 16).

After data pre-processing, we set the demarcate coefficient, the sort number and the confidence as 0.10, 3, 0.76 respectively. The gained result of land grading in Changqing is shown in figure 6. Furthermore, we adopted the same index system to grade the land in Changqing by the method of multi-factor synthetical evaluation, and it is also divided into 3 grades [16]. Meanwhile, we evaluated the price of commercial land in each town according to the samples of land market research and chose the price of commercial land in the highest level to validate the rationality of the results from the two methods. Figure 7 denotes the comparison of the results. We can see that, compared to the method of multi-factor synthetical evaluation, the grading result from spatial clustering is more practical and the grade borderline is more clear.

D. Decision tree analysis

Agricultural land classification can be seen as a classification assessment question about mixed spatial data which is from the quantized factors. And the result is the agricultural land grading. So we carry out the agricultural land classification, aiming to provide new method for agricultural land classifying.

The study area is Luanwan village in Pingyin county of Shandong in China, and the data resource is Pingyin agricultural land grading and classification database in 2002. There are 2353 assessing units. According to the characteristic of agricultural land resource and data resource, we choose the agricultural land grade index, location condition, cultivating advantage, land-use intension, land-use structure, land-use intensive degree and land operating benefit as test attributes. On the principle of equality of spatial distribution and sample' class, we select 2118 units (90%) as the training set while the other 10% as the testing set. The distribution of training-set samples in the study area is shown in figure 8.

Figure 8. Distribution of training-set samples in the study area

If the tree is too numerous and jumbled, some junior branches may be affected by abnormal value. So we need to prune it to avoid the over-fit. The optimal tree model of agricultural land grading in Luanwan is presented in figure 9. In decision tree analysis module, users can look over the information of each node. For example, in figure 9, it is the decision condition that if agricultural land grade index is between 3.5 and 2.5, location condition is below 2.5 and land operating benefit is below 1.5, then the grade is Ⅰ. This tree contains 22 "if-then" rules.

We utilized the 235 test samples to validate the assessing veracity of this decision tree model (the result is shown in figure 10). It was calculated that the assessing veracity rate accounts for 95.74%. Moreover, we evaluated the model by ten-iteration crossing validation and the average assessing veracity rate is 95.65%. Compared with the traditional methods, decision tree analyzing method not only can understand the spatial data better, find out the relationship between spatial data and non-spatial data and avoid the limitation of subjective estimation, but also is propitious to update agricultural land grading data fast.

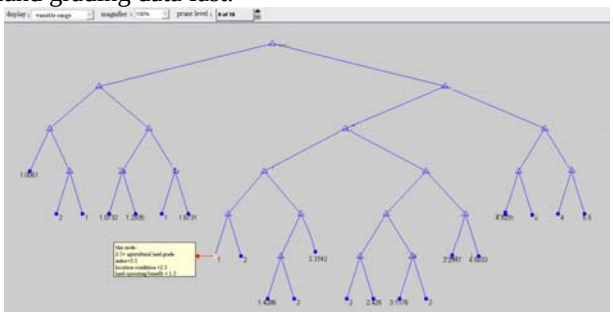

Figure 9. Decision tree model of agricultural land grading in Luanwan (after pruning)

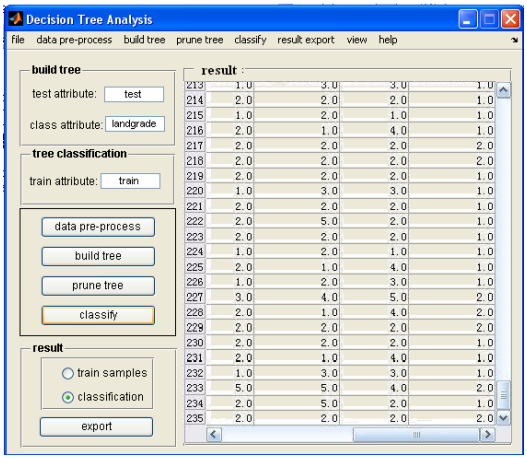

Figure 10. Decision tree classification result of agricultural land grading in Luanwan

Ⅴ. CONCLUSION AND SUGGESTION

This study has made an instructive discussion about constructing the spatial data mining system on the platform of MATLAB. In our further study, we should take more advantage of spatial reasoning and expert system, process SDM from object oriented spatial database, optimize the algorithm, improve the efficiency, and build an intelligentized spatial data mining system.

REFERENCES

© 2008 ACADEMY PUBLISHER

[1] Li Deren, Wang Shuliang, and Shi Wenzhong, "On Spatial Data Mining and Knowledge Discovery (SDMKD)" , *Geomatics and Information Science of Wuhan University*, vol.6, pp. 491–499, 2001.
[2] Zhang Ruiju, and Tao Huaxue, "The Study on Integration Between GIS and Spatial Data Mining", *Site Investigation Science and Technology*, vol.2, pp. 21–24, 2003.
[3] Zheng Xinqi, Dong Jinwei, and Zhong Buqing, "A spatial cluster method for prime farmland selection", *Geoinformatics 2007: Geospatial Information Technology and Applications.* 2007, p 67543J.
[4] Zhou Haiyan, "A Research on Spatial Data Mining", *Zhengzhou: Information Engineering University of PLA*, 2003.
[5] Liu Junqiang, Sun Xiaoying, Wang Xun, et al., "A New Algorithm for Mining Maximal Frequent Patterns", *Chinese Journal of Computers*, vol.10, pp. 1328–1333, 2004.
[6] Wang Shuliang, "View-angles of spatial data mining", *Journal of Tsinghua University, vol. (SUPPL.)*, pp. 1058~1063, 2006.
[7] Jia Zelu. "Design and realization of an intelligent land grading and evaluation information system", *Science of Surveying and Mapping*, vol.4, pp. 152–154, 2007.
[8] Wang Shengsheng, Liu Dayou, Xin Ying, et al., "Spatial reasoning based spatial data mining for precision agriculture", *APWeb 2006 International Workshops: XRA, IWSN, MEGA and ICSE*, pp. 506~510, Jan 16-18, 2006.
[9] Luo Jianjun, Yang Qi, MATLAB Tutorial, *Publishing House of Electronics Industry*, 2005.
[10] Kim YK, Lee SK, et al, "F-FDGPET in localization of frontal lobe epilepsy: comparison of visual and SPM analysis", *J Nucl Med*, vol.9, pp. 1167–1174, 2002.
[11] Ministry of Land and Resources, PRC. *Land use classification.* No.255, 2001.
[12] Zheng Xinqi, Zhao Lu, "Association rule analysis of spatial data mining based on MATLAB", *First international workshop on knowledge discovery and data mining, IEEE*, pp.76–80, 2008.
[13] Wang Yuejian, Liao Tiejun, and Huang Yun, "Study on urban land gradation——Jingyan County Sichuan Province as a case", *Territory & Natural Resources Study*, vol.1, pp. 15–16, 2006.
[14] Fan Li, Wu Qun, "A Study of Urban Land Grading in Jiangsu Province", *Scientific and Technological Management of Land and Resources*, vol.1, pp. 46–50, 2007.
[15] Wang Jing, Guo Xudong, "A Study of Scientific Regulation of Sustainable Land Use at County Scale in China", *Progress in Geography*, vol.3, pp. 216–222, 2002.
[16] Zhao Lu, Zheng Xinqi, "Multi-factor Synthetical Evaluation Method", *Journal of Anhui Agricultural Sciences*, vol.2, pp. 716–718, 2008.

Zhao Lu was born in 1985. She received B.S. degree and M.S. degree in Photography and Geographic Information System from Shandong Normal University, Ji'nan, China in 2006 and 2008, respectively.

She is currently pursuing Management science doctor's degree in Land Resource Management from China University of Geosciences, Beijing, China. She has authored/coauthored 14 papers in International/National journals and conferences.

Her current research interests include GIS development and application, land evaluation, spatial data mining and geographic calculation. Zhao Lu is member of China Society of Natural Resources (CSNR).

Zheng Xinqi was born in 1963. He received M.S. degree in Natural Geography from He'nan University, Zhengzhou, China in 1987. He received engineering doctor's degree in Graphics and Geographic Information System from Engineering Information University, Zhengzhou, China in 2004.

Currently he is professor in School of Land Science and Technology, China University of Geosciences, Beijing. He has authored/coauthored over 100 papers in International/National journals and conferences. He has host and participated almost 20 research subjects in national and provincial level. His current research interests include GIS, land evaluation and planning, spatial data mining, system simulation and geographical calculation.

Prof. Zheng is member of Key Laboratory of Land Regulation, Ministry of Land and Resources in China and the director of China Society of Natural Resources (CSNR).

Wang Shuqing was born in 1975. She obtained her B.S in Photogrammetry and Remote Sensing and M.S. in Graphics and Geographic Information System from Wuhan University, China in 1998 and 2004, respectively. She is now pursuing the engineering doctor's degree in China University of Geosciences, Beijing, China.

She is currently instructor in School of Land Science and Technology, China University of Geosciences, Beijing, China. She has authored/coauthored 4 papers in International/National journals and conferences. Her current research interests include GIS, RS image processing in land use.

© 2008 ACADEMY PUBLISHER