

GEOSPATIAL DATA MINING AND KNOWLEDGE DISCOVERY USING DECISION TREE ALGORITHM—A CASE STUDY OF SOIL DATA SET OF THE YELLOW RIVER DELTA

Jianting Zhang, Diansheng Guo, Qing Wan
LREIS, Institute of Geography, the Chinese Academy of Sciences
Beijing 100101, China
E-mail: {zhangjt, guods, wanq}@lreis.ac.cn

Abstract: This paper introduces the decision tree algorithm of machine learning methods from data mining and knowledge field and integrated it with GIS to provide a concrete example of improving geo-spatial analysis capability of GIS and mining geo-spatial knowledge from GIS database automatically and intelligently. In the first part of the paper, model structure of integration of GIS and decision tree algorithm, procedures of construction of decision tree rules and methods of calculating information entropy was demonstrated in detail. In the second part of this paper, we use the proposed method on the soil data set of Yellow River Delta (YRD). The data set consists of more than 600 polygons and four attributes, namely soil structure, soil essence, soil salinity and soil type. The association relationship between soil structure, soil essence, soil salinity and soil type was constructed using decision tree method. From the derived rules, it is easy to find their geo-spatial interpretation by domain experts. Finally based on our preliminary study and application of decision method, we draw the conclusion that decision tree algorithm is a good method for mining interpretable geo-spatial rules from GIS database and of great help to geo-science research, although lots of work still needs to be done.

Keywords: Decision Tree Method, Geo-Spatial Data Mining, Spatial Association, Yellow River Delta (YRD)

INTRODUCTION

GIS has gained extensive popularity in geo-spatial data management and its importance has been widely recognized [1]. However, lacking powerful analysis function has always been the bottleneck that restricts its wide application. Some researchers have tried to build GIS-based geo-expert system to enhance the analysis capability of GIS. But traditional symbolic expert system in geo-spatial analysis has the innate deficit in geo-spatial knowledge acquisition and update, thus it has many problems in practical implementation [2]. In recent years more and more concepts and methods of machine learning, data mining and knowledge discovery have been introduced into geo-spatial analysis. Geo-spatial data mining and knowledge discovery from huge volume of GIS database has lead geo-spatial data analysis to another whole new idea--knowledge acquisition through computation. Some researchers have put forward the concept and framework of spatial-knowledge discovery [3][4], the author himself also provides a literature overview applications of geo-spatial data analysis using machine learning algorithms and discuss some related problems of geo-spatial data mining and knowledge discovery [5]. But in general, essential work on this field is still lacking domestic.

Data mining algorithms mainly include traditional Probability and Statistic (PS), Artificial Neural Network (ANN) and Genetic Algorithm (GA), Decision Tree (DT), Rough Set (RS). Bayesian Probabilistic Network Learning (BN), etc. PS and ANN has gained more popularity than other algorithms in geo-spatial data analysis. However, the requirement that

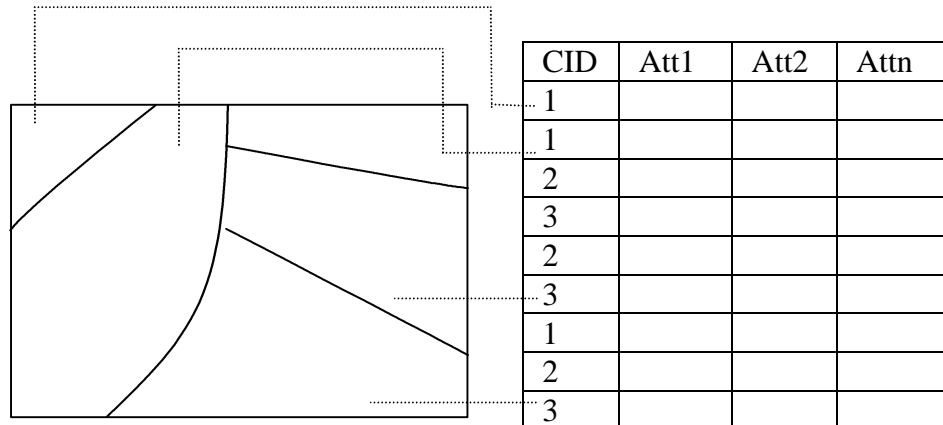
samples must have certain probabilistic distribution function of PS method is difficult to satisfy while ANN's connection weights can not suggest human-interpretable geo-spatial rules. Comparatively, by computing categorical information entropy, DT's merits of producing definite, apparent and human-interpretable rules have attracted more and more research interest. Some distinctive academic journals such as IJGIS [6], IJRS [7] and PE&RS [8] have published some research articles and reports on this topic. But most of the work focuses on remote sensing image classification and was not introduced into GIS.

This paper introduces the concepts and methods of decision tree algorithm, puts forward a model for the integration of GIS and the algorithm and then apply the algorithm to the soil data set of Yellow River Delta (YRD). The proposed algorithm built decision tree rules between soil essence, soil structure, soil salinity and soil classes. The experiment shows that DT has good prospects in geo-spatial data mining and knowledge discovery. As a general algorithm for categorical coverages, DT can be closely coupled into GIS and thus provides GIS a general powerful analytical tool.

DECISION TREE FOR MINING GEO-SPATIAL ASSOCIATION RULES IN GIS

Data Model

The basic model of geo-spatial association in GIS is shown as Fig. 1. In the attribute table of a coverage, we define one of the attributes as decision attribute (CID) and the others as condition attribute (Att1-Attn). Decision attribute must be discrete categories and condition attributes can either be discrete categories or continuous variables. If each patch in single coverage has multi-attributes, DT can be directly used to the data set. If we have multiple coverage and each of them have only one attribute; then we can use Overlay to produce



minimum patches of all the coverages. Each minimum patch will have multi attributes derived from the coverages and then the proposed method can also be used.

Fig. 1 General GIS Model of Geo-Spatial Association Using Decision Tree Method

Modelling Procedures

The basic procedures of building a decision tree is shown as fig.2.

- a) Calculating the information entropy of each partition or combination of each attribute using different method according to whether it belongs to discrete category or continuous variable. The method will be described in detail in the following section.

- b) Take the attribute which has the maximum information entropy reduction as the partition attribute and use the corresponding partition point or combination sets to divide the whole data set into two separate data sets and thus finish one partition.
- c) Repeat the above procedures for each data set until all the samples in the data set belong to a certain category (In this condition the information entropy in the data set equals zero and no partition is needed) or number of samples in the data set less than the predefined threshold (In this condition, the calculated information entropy has little statistical meaning).
- d) The final result will form a tree with root of the first partition attribute and the leaf of last partition and the stems of each partition condition.

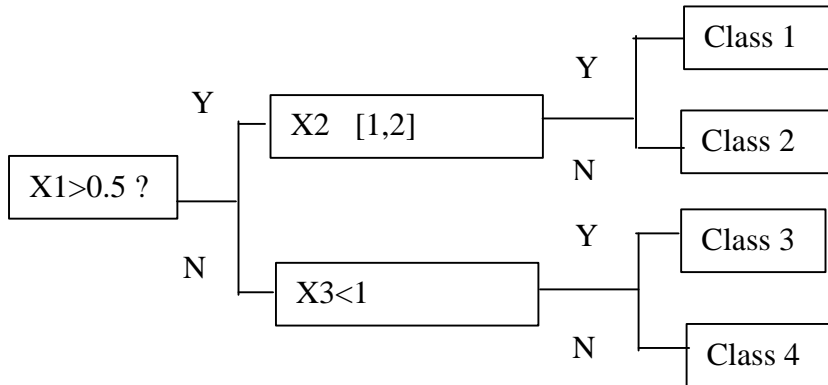


Fig. 2 Basic Procedure of Constructing Decision Tree

Calculation of Information Entropy

- e) Continuous Variable: first sort the whole data set according to the attribute value, then increase the partition value from the start point with a step value of ΔX , thus each partition value will be $X_i = X_0 + i * \Delta X$. Each X_i divides the whole data set into two sub data set. The equation for calculating the total information entropy of the partition is shown as formula 1.

$$dis = \sum_b \frac{n_b}{n_t} \sum_c - \frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b} \tag{1}$$

- f) Where n_t is the total area of the coverage, b is the number of partitioned sub sets (In this paper we adopt the method of bi-partition and thus $b=2$), c is the number of categories of decision attribute of each sub set, n_b is the total area of subset b and n_{bc} is the area of category c in subset b . The concrete example is graphically showed in the left side of Fig. 3
- g) Discrete attribute: try each combination of all the categories. A data set with n decision categories will have 2^n combinations. For each combination the equation for calculating information entropy is shown as formula 2. It has similar form to formula 1, but has different meaning.

$$dis = \sum_b \frac{n_b}{n_t} \sum_c - \frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b} \tag{2}$$

h) Where n_t is the total area of the coverage, $b=1$ and 2 indicates a combination set and its supplement set, c is the number of classes of decision attribute of the combination set or supplement set, n_b is the total area of subset b ($b=1,2$) and n_{bc} is the area of class c in subset b . The concrete example is graphically showed in the right side of Fig. 3

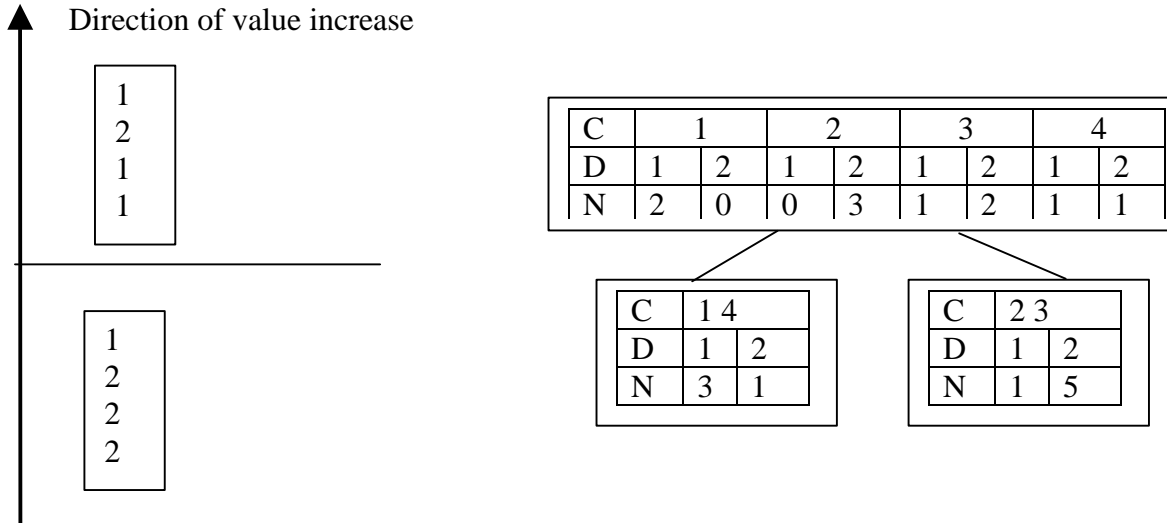


Fig 3 Split of Continuous and Discrete Variable in Constructing Decision Tree

EXAMPLE STUDY OF YRD SOIL DATA SET

In the soil data set of YRD, there are more than 600 polygonal patches and four attributes. The three condition attributes are soil structure (sl_str), soil essence (sf_ess) and soil salinity (salt) respectively. The decision attribute is soil type (t). Four thematic maps of each attribute are shown as fig. 4.

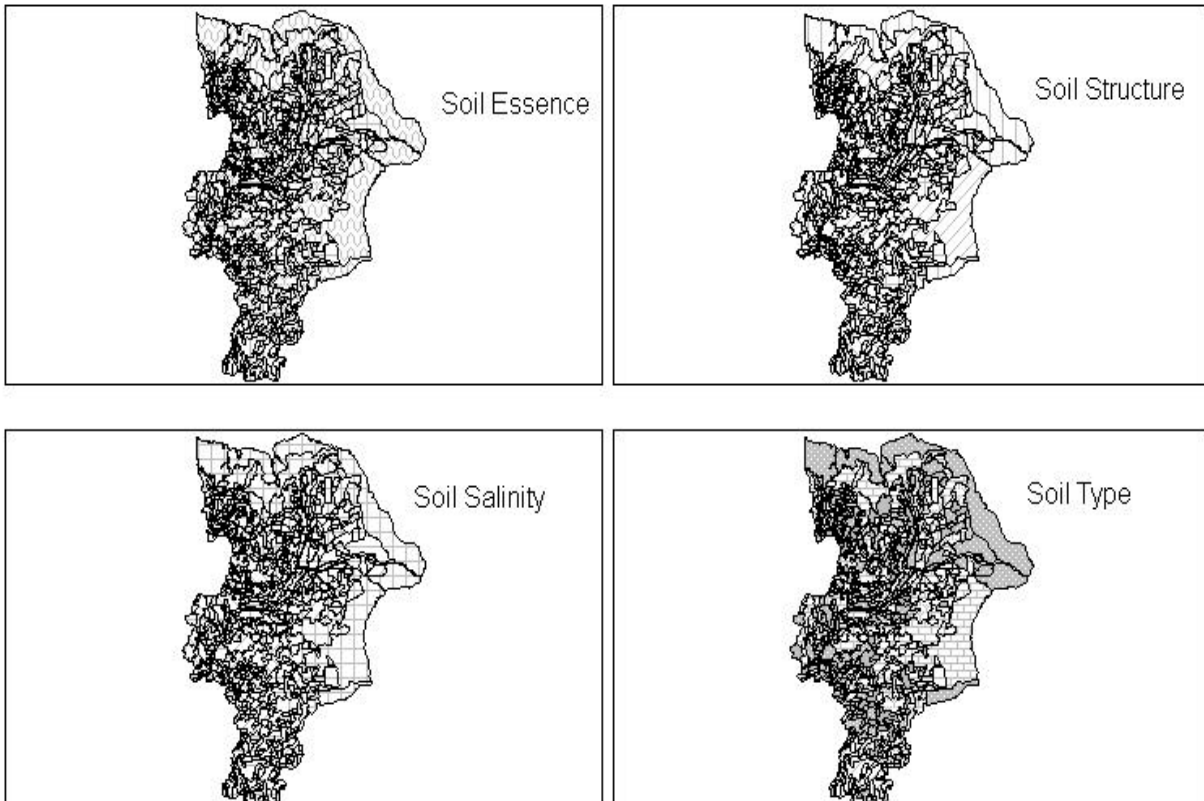


Fig 4 Maps of Soil Structure, Essence, Salinity and Type in YRD Soil Data Set

The decision tree rules derived from decision tree algorithm are shown as Fig. 5. Codes and their meaning are shown in Table 1.

```

salt in {3,4,5}: t8 (258.0)
salt in {1,2,6}:
| salt in {1,6}:
| | sl_str = 1: t1 (14.0)
| | sl_str in {8,9,10,13,14}: t9 (174.0)
| salt = 2:
| | sl_str in {8,9,10,11,12,13}: t5 (150.0/11.0)
| | sl_str in {1,2,3,4,5,6,7}:
| | | sl_str in {1,5,6,7}: t4 (9.0/1.0)
| | | sl_str in {2,3,4}:
| | | | sf_ess = 7: t3 (8.0)
| | | | sf_ess in {4,5,6}:
| | | | | sl_str = 2: t2 (5.0/1.0)
| | | | | sl_str in {3,4}: t3 (17.0/5.0)
    
```

Fig.5 Geo-spatial Association Rules of YRD Soil Data Set Using Decision Tree Method

Tab. 1. Attributes Codes and their Corresponding Meaning in YRD Soil Data Set

Soil Structure sl_str		Soil Essence sf_ess		Soil Salinity salt		Soil Type t	
1	Water	1	Water Area	1	Water Area	1	Water Area
2	Homogeneous	3	Sandy loam	2	<0.1%	2	Calcaric Combisols
3	Upper clay Layer	4	Sandy clay loam	3	0.1-0.2%	3	Gleyic Combisols
4	Deep clay Layer	5	Clayey loam	4	0.2-0.4%	4	Calcic Vertisols
5	Upper sand Layer	6	Silty clay	5	0.4-0.8%	5	Calcaric Fluvisols
6	Upper thick conglomerate Layer	7	Clay	6	>0.8%	6	Gleyic-Calcaric Fluvisols
7	Deep thick conglomerate layer			7	Saline-alkali land	7	Cambic-Calcaric Fluvisols
8	Isotropic sandy					8	Salic Fluvisols
9	Isotropic loam					9	Gleyic Solonchaks
10	Isotropic clay					10	Anhorosols
11	Argillaceous						

12	Agrillaceous bottom sand soil						
13	Arenaceous						
14	Interstratified clay						

By discarding rules with small number of samples or high percentage of exceptions, we can build the following attribute association rules:

- i) If soil salinity is greater than 0.8%, then soil type solely belongs to *Salic Fluvisols*.
- j) If soil salinity is between 0.1% and 0.8% then soil type solely belongs to *Gleyic Solonchaks*.
- k) If soil salinity is less than 0.1% and soil essence belongs to one of ... then soil type belongs to *Calcaric Fluvisols*.
- l) If soil salinity is less than 0.1% and soil essence belongs to one of ... and soil structure belongs to *Clay*, then soil type belongs to *Gleyic Combisols*.

Those decision tree rules reflect the clear geo-spatial patterns of YRD soil category classification. That is, soil in coastal area has the highest salinity and accordingly belongs to *Salic Fluvisols* and soil with less salinity belongs to *Gleyic Solonchaks*. Soil salinity is the main factor for soil classification with high salinity soil while soil essence and soil structure are main factors for the classification with low salinity soil. Such as soil salinity of *Gleyic Combisols* and *Calcaric Fluvisols* are all less than 0.1%, but soil essence of the former mainly belongs to .. and the latter belongs to

CONCLUSION AND FUTURE WORK

This paper introduces machine learning's decision tree algorithm from data mining field and integrates it into vector GIS and provides a concrete example of general analytical method for polygonal categorical coverage. From the work we have done we can draw the following conclusions:

- m) Polygonal thematic map is an important data type in GIS database and decision tree method provides an effective and efficient to deal with this kind of GIS data. The full integration of decision tree method with GIS will improve the analytical capability of GIS and provides an automatic and intelligent method for geo-spatial data analysis.
- n) Decision tree algorithm can produce explicit association rules of multiple attributes in the thematic coverage, they are easy for domain experts to interpret and examine. Compared with the ANN's blank-box model, DT model has more advantages in geo-spatial data analysis.
- o) Knowledge discovery from database is an extension for traditional geo-expert system. Rules generated by decision tree method and examined by domain experts can be used in the geo-expert system. It will greatly help Geo-ES in solving its bottleneck of knowledge acquisition.

This paper will be further studied in the following aspects:

- p) There are many methods for attribute partition and only information entropy reduction measurement was used in this paper. Comparison of the results of different methods is now undertaking and will be further studied.
- q) For continuous attributes, partition of data set into sub sets using information entropy method has the nature of axis paralleling. However it is not true in some circumstance and efforts should be paid to solve the problem.

- r) Decision tree method not only can be used to produce association rules but also can be used in classification using the established rules and thus make prediction. The next step of our work will be focused on these aspect and make comparisons with ISODATA and ANN, etc.

REFERENCES

- Chenghu Zhou, Quanqin Shao, On Application of Geographical Information System, *Acta Geographica Sinica, Supplement*, vol. 52, 1997
- Jianting Zhang, Youliang Qiu, Application of Artificial Intelligence and Expert System in Geo-Studies: An Overview, *Progress in Geography*, 16 3 1998, 44-51
- Deren Li, Tao Cheng, Knowledge Discovery From GIS Database, *Acta Geodaetica et Cartographica Sinica*, 24(1), 1995, 37-43,
- Jianting Zhang, Some Discussion about Geo-spatial Data Mining and Knowledge, Forthcoming, *Geography Research*.
- Kaichang Di, Deren Li, Deyi Li, A framework of Spatial Data Mining and Knowledge discovery, *Journal of Wuhan Technical University of Surveying and Mapping*, 22(4), 1997, 328-332
- M.Hansen, R.Dubayah and R.Defries, classification trees: an alternative to traditional land cover classifiers, *Int. J. Remote Sensing*, 1996, vol.17 No. 5. 1075-1081
- P.W.Eklund, S.D.Kirkby and A.Sallim, Data mining and soil salinity analysis, *Int. J. Geographical information science*, 1998, vol. 12, No. 3, 247-268
- A machine Learning approach to automated knowledge-base Building for Remote Sensing image Analysis with GIS data, Xueqiao Huang and John R.Jensen, *Photogrammetric engineering & Remote Sensing*, vol. 63, No. 10 , October 1997, 1185-1194