THE APPLICATION RESEARCH OF KNOWLEDGE DISCOVERY TECHNIQUES BASED ON ROUGH SET IN DECISION SUPPORT

WANG Shaohua [a], BIAN Fulin [a]

[a] Research Center of Spatial Information & Digital Engineering,Wuhan University, Hubei,P. R. China, (430079)
snoopywsh@163.com

Commission VI, WG II/5

KEY WORDS: GIS,Knowledge,Spatial,Rough set,Decision support

ABSTRACT

With the application and development of science and technology, the tremendous amount of spatial and nonspatial data have been stored in large spatial data bases. Analysing them for decision is badly in need of spatial data mining and knowledge discovery to provid knowledge. In recent years, some efforts in knowledge discovery have focused on applying the rough set method to knowledge discovery. In this paper, the application of knowledge discovery method based on rough set in land use decision support system is discussed. First the characteristic and development of knowledge discovery method based on rough set are briefly stated. Second, the characteristic of spatial data in GIS is discussed. Third, a knowledge discovery method based on rough set is put forward in land use decision support system. The procedures for this method, the algorithm and key matters are also analyzed. Finally, rules extracted by the method shows a good result. This method has solved the problem of obtaining the decision rule in DSS effectively.

1. INTRODUCTION

With the rapid development of the technology of data acquisition and database,extensive data in geographical information systems is increasing constantly. The present GIS systems mainly have such functions as data inputting, inquiry and statistics of those data,etc..Hence,their analysis functions is still very weak and not flexible, and cannot find relations and rules in the data effectively, so it is very difficult to extract the implicit mode to solve the problem on complicated spatial decision.

Data mining technology provides a new thoughts for organizing and managing tremendous spatial and nonspatial data. Rough set theory is one of the important method for knowledge discovery,which was firstly put forward by Pawlak in 1982. This method can analyze intactly data, obtain uncertain knowledge and offer an effective tool by reasoning.

GIS contains large amount of spatial and attribute data,and its information is more abundant and complicated than those stored in general relation databases and affair databases. The application of rough set theory to discover knowledge for dicision in spatial database is increasingly important in the construction of GIS system.

Taking the land use dicision support system as an example,this paper presents a method for knowledge discovery in spatial databases,on the basis of rough set theory, and illustrates its process.

2. OVERVIEW

2.1 GIS and Decision support

The geographical information system began to develop rapidly in 1960s. The most important characteristic of GIS technology is integrating and managing tremendous multi-subject spatial and attribute data.It can connects such attribute information as the society, economy, population, etc. with spatial position of the earth surface to establish a complete decision information database for inquiry, analyse and display. The rapid development of information technology and the new requirments in this field has revealed the defects of affair-oriented GIS which is urged to tranfer from management to dicision support.

An important trend of GIS development is intelligent DSS s(IDSS)(Li Deren,1995).GIS are increasingly being used for decision-making, yet it is still not enough to solve semi- or ill-structured decision problems that own the character of fuzziness and uncertainty. This makes the study on knowledge-based GIS interesting to researchers (Cohena & Shoshany, 2002; Yamadaa, et al., 2003). Knowledge is the foundmental of DSS,and many researchs are being focused on discoverying knowledge from tremendous database.

2.2 Spatial Data Mining

Spatial data mining (also called geographical knowledge dicovery),which is a branch of data mining, puts emphasis on extracting implicit knowledge,spatial relations and other significative modes from spatial data.

Different from general mining tasks,spatial data mining is mainly involved in the researches on the probability distribution modes, clustering and classification characteristics,reliance relation between attributes etc. of spatial data. It is much complicated than general data mining in relation database,which is shown in two aspects.

- Huge amount of spatial data and the complexity of spatial data type and spatial visit method. That is,besides the nonspatial information such as word,characters, spatial data,contains the spatial information such as topological relations, distance relation and direction relation.
- The relations between spatial data are connatural.The relation between spatial entities is connatural,so is the relation between such attributes as population,economy and social development in spatial entities,which makes the spatial data mining more difficult.

Rough set,consisting of upper approxmate set and lower approxmate set, is a tool for intelligent dicision analysis dealing

with inaccurate, uncertain and incomplete information.It is suitable for spatial data mining based on the uncertainty of attributes and provide a new approach for GIS attribute analysis and knowledge discovery.In spatial data mining,the application of rough set can analyze the importance, uncertainty, consistency and dependent index of attributes,study the effect of attributes' dependent index on decision-making, reduce the data, attribute table and attributes' dependent index, discovery the relativity of data, evaluate the absolute and relative uncertainty of data, and obtain causality in the data, produce minimum decision and classification algorithm, etc.

3. ROUGH SET

3.1 Basic conception

Assume a domain set U of a target,let $X \subseteq U$ and R indicates an equivalent relation,When X is the ombination of some basic categorieses of R ,then X can be defined by R;otherwise X can not be defined by R. The sets which can be defined by R are subsets of the domain set, which are called R accurate sets, and can be precisely defined in the knowledge base K.On the contrary, the sets which can not be defined by R can not be defined in the knowledge base, which are called R rough sets. Routh set can be described by two accurate sets and a boudery set:

Lower Approxmate Set of X on R is definded as:

$$R_{-}(X) = \bigcup \{Y \in U / R \mid Y \subseteq X\}$$

Upper Approxmate Set of X on R is definded as:

$$R^{-}(X) = \bigcup \{Y \in U / R \mid Y \cap X \neq \Phi\}$$

Boundary Set of X on R is definded as:

$$bn_R(X) = R^{-}(X) - R_{-}(X)$$

posR(X)=R_(X) is defined as positive domain

while negR(X)=U-R_(X) is defined as negative domain

3.2 Knowledge representation, Reduction, and Core

Knowledge representation is achieved through knowledge expression system. Its basic composition are object sets whose knowledge is described by the attributes of targets and themselves.

A knowledge representation system can be expressed by

$$S = <U, C, D, V, F>$$

where U is the domain set.

$C \subseteq U$ is attribute set , $C = \{a_1, a_2, \ldots, a_m\}$ is the condition attribute set (note should be taken that C contains spatial constraint conditions), $D = \{d_1, d_2, \ldots, d_x\}$ is the decision attribute set,

V is the field collection composed of $C \cup U$, viz. $V = \bigcup_{p \in A} V_p$, V_p is the field of attribute p, f is an information function, viz. $f: U \times A \rightarrow V$,

Let attribute set:

$$B = \{b_1, b_2, b_3 \ldots, b_m\} \subseteq A \Rightarrow V_B = V_{b1} \times V_{b2} \times V_{b3} \times \ldots \times V_{bm}$$

Define the map F_B

$U \rightarrow V_B$ to represent attribute value of field B.

R upper set of condition set C about domain set U can be expressed as: U/R_C

R upper set of decision set D about domain set U can be expressed as: U/R_D

Define U/R_B as the equivalent of a transaction, then U/R_C is the transaction of condition, U/R_D is the transaction of decision.

Then the upper approximate about the condition set for decision transaction is:

$$C^{-}(D_j) = \{C_j \mid C_j \in U / R_c \cap C_j \cap D_j \neq \Phi\}$$

Then the lower approximate about the condition set for decision transaction is:

$$C_{-}(D_j) = \{C_j \mid C_j \in U / R_c \cap C_j \subseteq D_j\}$$

Let the two sets G and R, r is an equivalent relation in R, g is an equivalent relation in G, if $pos_{R-\{r\}}(G) = pos_R(G)$,Then r in G is omissible.

If no element in R can be omissible, then R is independent.

Let $H \subseteq R$ H is independent, if $pos_H(G) = pos_R(G)$ then H is the reduction of G by R. From the definition, the lower approximate of G about H and R is the same, which maintains the classification of R and G. The all intersection of the relation in R forms the core of R and marked as core(R),viz.core$(R) = \bigcap red(R)$. The attributes in core set is the key factors that affect classifation based on R.

3.3 Dependent index of Decision Transaction

C_i is the condition of U/R_C,and D_j is the decision of U/R_D,let decision transaction based on condition transaction can be mapped as $CF_{ij}(C_i \rightarrow D_j)$ and $CF_{ij} = card(C_i \cap D_j)/card(C_i)$, if condition transaction C_j is belonged in the lower approximate $C_{-}(D_j)$ if decision transaction , $CF_{ij}=1$;otherwize if condition transaction C_j is belonged in in $U - C^{-}(D_j)$,$CF_{ij}=0$.

4. A KDD MOTHED BASED ON ROUGH SET

4.1 Spatial Object Information Tables

The knowledge representation system describes the domain set as a two-dimentional table in which each row indicates an object and each column indicates an attribute.Here,the attributes can be divided into condition attribute and decision-maiking attribute. In the process of knowledge discovery, condition attribute should be reduced first to remove repeated rows,then redundant attribute in each decision-making should be reduced.To reach the minimum decision rules in application, we should select effective attributes to indicates the domain set properly or approximately.

4.2 Minimum rule generation algorithm based on Rough Set

Generally, decision-makers have priori knowledge to the weight of every condition attribute. The weight is used for weighing the relative importance of attribute. In various decision-making, the same attributes may have different influence on decision-making, namely the weight is sensitivfe to the decision-making environment. The dependent index of attribute expresses the influence of the attribute on decision rule under present data environment, but can not reflect the decision-maker's priori knowledge. So, it is a comparatively reasonable method to combine them to select effective attribute. The processes are described as follows.

1. Propose two-dimentional data view, i.e. decision rule table, composed of condition attribute and decision attribute in the domain set;
2. Determine the data classification standard, express the attribute values in a standardized way, and remove unnecessary attributes;
3. If the decision rules of the knowledge expression system are exclusive, we can classify it into two sub-tables: one is an inclusive decision table; the other is an exclusive decision table. The latter is a kind of knowledge which can not be extracted from the present information, so we just deal with the former.

4. Calculate the dependent index of every attribute. The dependent index of every sub attribute can be obtained from the determined conditions. Certainly, the importance of attribute a can be expressed by analyzing the quotient of $pos_{B-\{a\}}(C)$ and $pos_B(C)$;
5. When removing the attribute whose dependent index is *0*, the positive domain of *U/C* is not affected. Therefore, according to the order of the priori weight, remove the attribute whose dependent index is 0 and whose priori weight is minimum.
6. Calculate the core of every decision rule and its possible reduction forms.
7. Select the attribute reduction table of effective decision rule according to certain principles and obtain the minimum rule.

In practice, there may exist a very big rule set in step *7*. Except for the specific cases of decision-making,such a big set is troublesome in practice. So, the most effective subset of attribute should be considered to correctly reduce or approximatly express the domain set. Generally, we judging a target, people would first take the atrribute with largest weight according to their priori knowledge into consideration. Thereby,we should select the reduction rules of the attributes with lager weights to represent the decision rule of the domain set. The following is a practical and effective method.

Assuming that the reduced decision attribute set is $\{a_1,a_2,\ldots a_m\}$,their priori weight are $p(a_1),p(a_2),\ldots p(a_m)$, respectively,and rule i probably has k reduction form, then the definition of the weight of each form is:

$$P_j = \sum_{j=1}^{m} (O(a_j) \times p(a_j))$$

Here,if a_j is an appointed value, then $O(a_j)$=1; if not an appointed value, then $O(a_j)$=0. Last, combing the reduction forms with lagest weight to obtain a practical and effective decision rule set.

5. CASE STUDY

Taking the land use decision support system as an example, we discuss the problem on the proper types of crops in a certain type of soil.

In Table 1, c_1, c_2, c_3, c_4 are the condition attributes; d is the decision-making attribute.c_1 indicates the elevation,c_2 soil type, c_3 crop type; c_4 annual average temperature and d output.

U	c_1	c_2	c_3	c_4	d
1	50	Red	Rice	25	Few
2	10	Brown	Wheat	12	Few
3	240	Red	Wheat	15	Few
4	320	Brown	Wheat	13	Lot
5	400	Black	Sorghum	5	Few
6	900	Brown	Rice	26	Few
7	600	Brown	Wheat	18	Lot
8	1250	Brown	Rice	22	Lot
9	1300	Black	Wheat	11	Few
10	1400	Red	Rice	21	Few

Table 1. Spatial Object Information Tables

Standardize the above table according to the classification standard:

Elevation□0□[0□100]□1□[100□500]□2□[500□1200]□3□[1200□∞]□

soil type□0□Red□1□Brown□2□Black□

Crop type□0□Rice□1□Wheat□2□Sorghum□

Annual average temperature□0□[-10□10]□1□[10□20]□2□[20□∞]□

Output□0□Few□1□Lot□

Then we have Table 2.

U	c_1	c_2	c_3	c_4	d
1	0	0	0	2	0
2	0	1	1	1	0
3	1	0	1	1	0
4	1	1	1	1	1
5	1	2	2	0	0
6	2	1	0	2	0
7	2	1	1	1	1
8	3	1	0	2	1
9	3	2	1	1	0
10	3	0	0	2	0

Table 2. Spatial information table after standardization

The weights of the attributes are listed as

c_1=0.35, c_2=0.3, c_3=0.2, c_4=0.15□

Analyzing the attribute one by one,we obtain the dependent index.

Let $C=\{c_1,c_2,c_3,c_4\}$,$D=\{d\}$, then the dependent index of D on C $CF=card(C\cap D)/card(C)$=1. We can see that the data views are inclusive.To $c1$,the dependent index of D on $c1$ is $CF_{c1}=card(C_{c1} \cap D)/card(C_{c1})$=5/8.Similarly,$CF_{c2}$=1/2,$CF_{c3}$=0,$CF_{c4}$=0.

According to the weights of the attributes,we can conclude that weight of attribute c_3 is larger than that of c_4. Therefore, c_3 is remained while c_4 is removed. In the data view without c_4, we can find that the dependent index of each attribute is greater than 0. So, each item can not be omitted. However, to obtain the reduced decision rule,we need to remove the unnecessary conditions in every decision rule, namely calculating the core of each rule.

To decision rule 1,

$F=\{[1]_{c1},[1]_{c2},[1]_{c3}\}=\{\{1,2\},\{1,3,10\},\{1,6,8,10\}\}$,

i.e. $[1]_{\{c1,c2,c3\}}=\{1\}$, $[1]_d=\{1,2,3,5,6,9,10\}$.

To find the unnecessary attributes of dicision rule 1,we should check whether the intersection of other attributes' subset is within the decision attribute's sub-set $[1]_d$.

$[1]_{c1}\cap[1]_{c2}=\{1\}$, $[1]_{c1}\cap[1]_{c3}=\{1\}$, $[1]_{c2}\cap[1]_{c3}=\{1,10\}$,

Then we can find the core of decision rule 1 is empty,which can be expressed in three forms: $c_1(1)=0$ and $c_2(1)=0$,$c_2(1)=0$ and $c_3(1)=0$, $c_1(1)=0$ and $c_3(1)=0$.

Similarly, we can obtain the core of each rule and its reduced form, listed in Table 3 and Table 4.

U	c_1	c_2	c_3	d
1	X	X	X	0
2	0	X	X	0
3	X	0	X	0
4	1	1	X	1
5	X	X	X	0
6	X	X	X	0
7	2	X	X	1
8	X	1	X	1
9	X	X	X	0
10	X	X	X	0

Table 3. Reduced spatial information table

U	c_1	c_2	c_3	d
1_1	X	0	0	0
1_2	0	X	0	0
1_3	0	0	X	0

2_1	0	X	1	0
2_2	0	1	X	0
3_1	X	0	1	0
3_2	1	0	X	0
4	1	1	X	1
5_1	X	2	2	0
5_2	1	X	2	0
5_3	1	2	X	0
6_1	X	1	0	0
6_2	2	X	0	0
7	2	X	1	1
8	3	1	X	1
9_1	X	2	1	0
9_2	3	X	1	0
9_3	3	2	X	0
10_1	X	0	0	0
10_2	3	X	0	0
10_3	3	0	X	0

Table 4. Expanded spatial information table

It can be seen from Table 4 that decision rules 4, 7 and 8 have only one reduction form, that decision rules 2, 3 and 6 have two reduction forms, and that decision rules 1, 5, 9 and 10 have three reduction forms. Thus,the knowledge expression system has $(1\times1\times1)\times(2\times2\times2)\times(3\times3\times3\times3)=648$ reduction forms.
According to the practical and effective principles, the largest weights of the rules are $1_3\square2_2\square3_2\square4\square5_3\square6_2\square7\square8\square9_3\square10_3$, respectively. Then, we can obtain the reduced practical decision rule as follows.
$c_1(0)c_2(0)\square c_1(0)c_2(1)\square c_1(1)\ c_2(0)\square c_1(1)\ c_2(2)\square c_1(2)\ c_3(0)\square c_1(3)\ c_2(2)\square c_1(3)\ c_2(0)\rightarrow 0$
$c_1(1)\ c_2(1)\square c_1(3)\ c_2(1)\square c_1(2)\ c_3(1)\rightarrow 1$

6. CONCLUSIONS

Rough Set Theory has been widely used in KD(knowledge discovery) since it was put forward. Having important functions in the expression, study, conclusion and etc. of the uncertain knowledge, it is a powerful tool which sets up the intelligent decision system. Actually, many knowledge systems are so rough that they make an obvious delay in the response time of an Intellectual Processing System by being put into a knowledge database directly. So, it is necessary to refine the knowledge which is extracted further. This article discussed how to express knowledge within an Information System with conditions-decisions forms in DSS, to get the potentialy reduced rules of decision-making by using Rough Set, spatial information tables on the basis of this and combining with analysis and reasoning with the priori knowledge from decision-makers, then to obtain a group of reasonable decision-rule sets by using the practical and effective principles, and finally to solve the problems of obtaining the decision-rules in DSS.

REFERENCES

Pawlak,Z.Rough Sets.International Journal of Information and Computer Science, 1982, 11

Pawlak, Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic, 1991.

Lin, T.Y., and Cercone, N., Eds. Rough Sets and Data Mining. Kluwer Academic. 1997.

Mrozek, A., and Plonka, L. Rough sets in industrial applications. In Rough Sets in Knowledge Discovery, Vol. 2. L. Polkowski and A. Skowron, Eds. Physica Verlag, 1998.

Application of a spatial decision support system in managing the protection forests of Vienna for sustained yield of water resources Forest Ecology and Management Volume: 143, Issue: 1-3, April 1, 2001, pp. 65-76 Vacik, Harald; Lexer, Manfred J.

Lawrence A. West Jr., Traci J. Hess. Metadata as a knowledge management tool: supporting intelligent agent and end user access to spatial data. Decision Support Systems 32 (2002) 247– 264

Massimo Dragan, Enrico Feoli, Michele Fernetti, etc.. Application of a spatial decision support system (SDSS) to reduce soil erosion in northern Ethiopia. Environmental Modelling & Software 18 (2003) 861–868

Using a spatial decision support system for solving the vehicle routing problem Information and Management Volume: 39, Issue: 5, March, 2002, pp. 359-375 Tarantilis, C.D.; Kiranoudis, C.T.

Yafit Cohena,Maxim Shoshany. A national knowledge-based crop recognition in Mediterranean environment. International Journal of Applied Earth Observation and Geoinformation 4 (2002) 75–87

Agrawal R , Imielinski T , Swami A. Mining Association Rules between Sets of Items in Large Databases. The1993 ACM-SIGMOD ,Washington ,1993

Ziarko W. Introduction to the Special Issue on Rough Sets and Knowledge Discovery. International Journal of Computational Intelligence ,1995 ,11(2) :223□226

Lavingto S,Dewhurst N,Wilkins E, et al. Interfacing Knowledge Discovery Algorithms to Large Database Management Systems. Information and Software Technology ,1999 ,41(9) :605□617

Fayyad U M , Piatetsky- Shapiro G,Smyth P. From Data Mining to Knowledge Discovery:An Overview ,Advances in Knowledge Discovery and Data Mining. In : Fayyad UM. eds. Advances in Knowledge Discovery and Data Mining. Massechusetts :AAAI/ MIT Press ,1996. 1□36

Slowinski R. Intelligent Decision Support : Handbook of Applications and Advances of Rough Set Theory.Netherland : Kluwer Academic Publishers ,1992. 1□5

Duentsh I ,Gediga G. Statistical Evaluation of Rough Set Dependency Analysis. International Journal of Human-Computer Studies,1997(46) :589□604