COMPREHENSIVE KNOWLEDGE DISCOVERY: THEORY, CONCEPT AND APPLICATION

Sha Zongyao [a, *], Bian Fuling [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, 430079, WuHan, China - (zongyaosha, Fuling Bian)@263.net

KEY WORDS: comprehensive knowledge discovery, knowledge discovery algorithm, spatial association rule, knowledge expression system, data mining

ABSTRACT:

The principle of comprehensive knowledge discovery is proposed in this article. Unlike most of the current knowledge discovery methods, comprehensive knowledge discovery considers both the spatial relations and attributes of spatial entities or objects. We first introduce the theory if spatial knowledge expression system and some concepts that are the base for our research. Those concepts include: spatial object classification, spatial relations, comprehensive knowledge discovery, comprehensive knowledge and SUIT, etc.. In theory, SUIT records all information contained in the study objects, but in reality because of the complexity and varieties of spatial relations, only those factors that have interest to us are selected. The selected factors constitute the to-be processed data which are a subset of SUIT. In this study, we select spatial association relation as the research emphasis. In order to find out the comprehensive knowledge from spatial databases, an efficient comprehensive knowledge discovery algorithm called recycled algorithm (RAR) is suggested.

As an example, we give a case study based on our proposed algorithm to get comprehensive knowledge. The research areas are agricultural land in two adjacent counties in northern China. The climate of northern China is very in lack of rain and only those crops that accustomed to arid environment can survive. By RAR, spatial association rules can be mined and used to make decision on crop planting distribution in agricultural planning.

1. 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 existing 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 (Leung Yee, etc., 2001; W. Lu, etc., 1993; KERRY TAYLOR, 1999). According to knowledge types, it can be grouped into spatial information generalization, spatial association, spatial classification and spatial clustering. In data mining research area, there exist two trends: mining attribute data mining omitting spatial character and mining spatial relation omitting attribute associated with spatial objects. Both of those trends are not complete in expressing the real world. Purely attribute data miner ignores the fact that 80% of the earth data surrounding us has spatial character while purely spatial data miner usually emphases less on the importance of spatial object attribute which are deep and full description of an object. Both experience and theory study prove that knowledge discovery from spatial databases (KDSD) cannot be independent of spatial objects attributes and should be associated together to give a complete expression of spatial objects. The association rule discovery problem in particular has been widely studied and has been the focus of many studies in the last few years and spatial association rule mining has become one important aspect of KDSD (S Rahayana and A Siberschatz, 1998). General transaction association rule mining cannot explore the implicit spatial rule in database, so method that can integrate mining both spatial and attribute features is urgently needed. Since the spatial relations are complex and are not easy to express, it is very important to construct a suitable spatial data mining model which can ease the process of KDSD. We present in this paper an approach to integrate mining spatial relation and the attribute character of spatial objects from large spatial data repositories. We also use this approach to explore the rules that lies behind large database collected from a case study area. The rules mined proved to be valuable and understandable.

2. THEORY AND CONCEPTS

2.1 Spatial Knowledge Expression System

Let $S=(U, C, D, V, f)$, and $U=\{u_1, u_2, \cdots, u_n\}$. U is a finite set of objects, $A=C \cup U$ is attribute set , $C=\{a_1, a_2, \cdots, a_m\}$ is the condition attribute set (note should be taken that C contains spatial constraint conditions), $D=\{d_1, d_2, \cdots, d_x\}$ is the decision attribute set, V is the field set composed of $C \cup U$, viz. $V=\cup_{p \in A} V_p$, V_p is the field of attribute p, f is an information function, viz. $f: U \times A \rightarrow V$. S is defined as formalization definition of spatial knowledge expression system (or SKES).

In the view of the form of SKES, there is no difference between SKES and the general knowledge system often outlined in artificial intelligence. However, the condition attribute set of SKES includes both spatial and attribute constraints. As for spatial constraints, different spatial relation type may have different forms. For example, if we consider spatial clustering, a spatial object may be given a constraint that it must be a certain clustered group. At the same time if we research on spatial association, we should first classify the spatial objects (into n categories) and then construct an attribute set with n-dimensions. The value of each object in the n-dimensions set will be given spatial index value (fuzzy index type) or 0-1 (Boolean index type) according to the spatial association we intend to include in a research project.

Spatial knowledge data mining aims at mining some interesting patterns that are unknown before analysis. Because of the complexity of spatial database, which not only contains attribute data but also spatial relations (topology relation, metric relation, orientation relation, etc.), mining spatial knowledge imposes more challenges. Currently, the main research area of spatial knowledge data mining is basic theories, optimised algorithms and applications (Bian Fuling, etc., 2001; Ziarko W, 1995). SKES is intend to abstract irregular data from the real world that contains valuable information and simplify the disposing processes.

2.2 Some Concepts

Spatial object classification: Classification assumes that homogeneity and heterogeneity exist between objects. The standard of classification is the essential factors of objects that can be used to identify a given object. According to such standard, spatial objects then can be grouped into a several divisions. The research of spatial relations is undergone based on the frame of spatial classification. Without classification, all spatial objects are referred to be one thing and the relation does not exist. If we say object A and B have an association relation, we know A and B do not belong to the same group.

Spatial relations: Researching spatial relations is a key area in GIS theory and application, and an important function of GIS is embodied by spatial analysis (Sauchyn, DJ, Yong Xongchao, 1991; T.Q. Zeng and Q. Zhou, 2001; Zhang T, etc., 1997). The footstone of spatial analysis is to understand spatial relations. The methods for describing spatial relations include intersection-based model, interaction-based model and hybrid method based on *voronoi* graph. According to semantically relation, spatial relations can be divided into topological relation, ordinal relation and metric relation, etc.. But in reality, only one or two is selected to research on.

Comprehensive knowledge discovery: Comprehensive knowledge discovery is to analyse comprehensively on spatial character as well as attributes of spatial entities and to find out deep regulations that are stored implicitly in attributes information and spatial information of research objects. For example, in the process of analysing on influential factors upon crop yield, we usually only consider possible attribute factors such as climate, soil fertility, soil texture, etc. but ignore spatial information (climate and soil fertility distribution) that may contain spatial association patterns. Those unknown patterns can then be used to support decision for crop planting area planning and yield evaluation.

Comprehensive knowledge: This is referred to be the rules that are found out by the method of comprehensive knowledge discovery. The patterns, containing both generalized spatial and attribute information, are understandable and having potential applications.

Spatial union information table (or SUIT): The above-mentioned spatial information includes graphical information, topological information and attribute information of spatial entities. SUIT is defined as information table containing graphical information, topological information and attribute information of spatial entities. This table can be separated into two parts with spatial relations (SR) that record classification and relations of spatial entities and attribute information (AR) which records attribute fields of spatial entities. In a formalized form, SUIT is expressed as SUIT（T，SR，SRV, AI，AIV）, while T stands for the whole set of spatial entities, and SRV and AIV are index value or representative mode of spatial relations and attribute value of spatial entities. It is possible to obtain all possible valuable information (such as spatial information generalization, spatial association rules, spatial classification and clustering, etc.) by processing on SUIT. In practice, it is impossible to consider all factors. For specific purposes, some simplifications have to be made. Let SUIT$'$ as the actual study goal, and T$'$, SR$'$, and AI' are subsets of spatial objects, spatial relation objects and attribute of spatial objects respectively, viz. T$'\subseteq$ T, SR$'\subseteq$ SR, AI'$\subseteq$AI. As shown in Figure 1, the sub-sets of spatial objects: T$'$={A，B，C，D，A}, the classification of spatial objects is: SR$'$ ={A，B，C，D}, suppose attribute field sets AI'={area，perimeter}, the content of SUIT$'$ is shown in Table 1. The value of the elements in Table 1 shows the relative neighbourhood index between entities shown in Figure 1. In the following sections, we will use SUIT to stand for SUIT'.

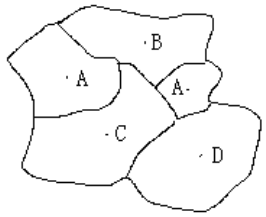

Figure 1. Diagram showing spatial object relation

T	A	B	C	D	area	peri
A	0	0.5	0.7	0	1.95	2.74
B	0.9	0	0.3	0	1.90	2.88
C	1.1	0.3	0	0.6	2.13	3.11
D	0.5	0	0.6	0	1.89	2.69
A	0	0.4	0.4	0.5	0.84	1.39

Table 1. SUIT$'$ for spatial objects

Figure 1 shows the fact of spatial adjacency relation of spatially neighbouring objects. Possible conditions other than adjacency relation can also be integrated in SUIT. The difference between spatial adjacency relation and other relations lies in the different application meaning of elements in Table 1. We are now researching on a method called influential field to qualitatively represent spatial relations by extending adjacency relation. In this method, spatial objects do not need share edge (in Figure 1 spatial objects share edge). They can be crossing, separating or adjacency. Figure 2 shows two separating spatial objects O1 and O2, and their influential fields. Each object in a given area is influenced by numerous fields caused by other objects. Those spatial objects can also be classified as A, B, etc. so for any object, its relation index value with other objects can be calculated according to the influential field model. By this way the possible relations between spatial objects can be extended.

3. OVERVIEW OF ASSOCIATION RULE AND SPATIAL ASSOCIATION RULE MINING

3.1 Association Rule and Spatial Association Rule

Association rule describes item relations in a database. In a mathematic language, let I={i_1, i_2, …, i_m}and it is an itemset called dataset, let D is a collection of all possible itemset.

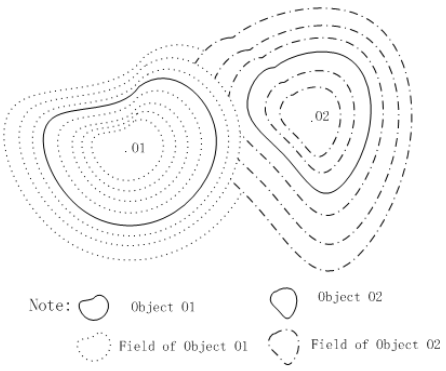

Figure 2. Influential field for two separating spatial objects

Transaction T is a sub set of I, vis. T⊂I, and every transaction is identified by TID. For dataset X, we call T includes X if and only if X⊆T. Association rule is usually presented in the shape like: X=>Y, here X⊆I, Y⊆I, and X∩Y＝Ø. X is called the condition of the rule while Y is result of the condition X. The confidence level for the rule X=>Y on the base of set D is defined as c% which means that out of all transactions in D there are c% records include both X and Y, while the support level of rule X=>Y is defined as s% that means there are s% out of all transaction records include X□Y. Confidence reflects intensity of rule and support reflects frequency of a rule. Rules whose support level are higher than the predefined support level are called frequent rules and both confidence level and support level are higher than predefined ones are called intensive rules. So we can see association rule represents relations between objects in macro view. This is different from association relation between two items or two spatial objects that represents a certain relation between items or objects in micro view.

The focus of spatial association rule mining is on spatial information. In formalized expression, it is described as: for spatial objects A and B (A, B do not belong to the same type) and complete objects set U, let R be the spatial relation term, A*R*B if A and B have an association relation R. For example, if along the sides of river A in 5 kilometres arrange area 80% of all area is distributed by agricultural fields then we say agricultural fields and river have a certain association relation. There exists a large difference in association mining between spatial database and transaction database. Firstly, spatial association rule in spatial database is more difficult to be mined because spatial data structure is usually unstructured. So data miners have to be familiar with a particular data structure and take primary data preparation to translate data structure into another one that is convenient to be handled. Secondly, spatial association relation is fuzzy. Items in transaction database have either associate relation or not, often referenced as 0-1 algebras. There is no middle state. For spatial database, however, association relation is co-determined by objects that belong to one or several types. Suppose object X, grouped as type O, has relation objects X1, X2, X3 and X4, and X1, X2 belong to type A, X2 and X3 belong to type B and type C respectively. As for object X, its contribution to spatial relation between type O and type A is the sum of relation index value of X1 and of X2. A predefined standard has to be set to determine whether object X is spatially associated with X1 (and X2).

3.2 Review of Algorithms in Mining Association Rules

The general method for mining association rule include the following procedures: (1)to find out all frequent itemsets; (2)to form association rules from frequent itemsets. For the given full item set U, if A⊆U is a sub set of U and sup(A)/ sup(U)>*Confidence*, where sup(X), *Confidence* stand for support level and confidence level for spatial type X, then the rule A=>U-A can be induced. In the above two steps, the first step is central. Once frequent itemsets have been got, it will be easy to form rules.

Classic association rule mining algorithms such as Apriori and DHP (Direct Hashing and Pruning), etc. are usually used to draw rules from transaction databases. The principle of Apriori is to generate large candidate itemsets by scanning database and calculate the happening times of each candidate itemset. One-dimensioned large itemset L_1 extracted from the large candidate itemsets. Next is to generate two-dimensioned large itemset L_2 based on L_1 and the database. Following the same method, n-dimensioned large itemset L_N can be formed and the n+1 dimensioned large itemset no longer exist. Sequential large itemsets $\{L_1, L_2, \ldots, L_N\}$ can be got. Because Apriori is very time consuming in generating large itemsets Park proposed hashing pruning called DHP algorithm.

Besides the above mentioned algorithms, literature (Agrawal R, etc., 1993; Lavingto N, etc., 1999) also proposed generalized association rule, multi-level association rules, quantitative association rules mining respectively. But almost all of those algorithms need scanning database many times, which greatly reduce their efficiency.

All of those methods are directed to transaction database, however. As for spatial association rule mining, it is also possible to apply after a little modification. But the spatial database must be based on suitable data model. Even so, the efficiency is low. So in section 5, we will give an efficient algorithm (RAR) to find spatial association rules. As preparation, neighbourhood index value and the generation of SUIT are in the field of data model construction.

4. INDEX VALUE CALCULATION FOR SPATIAL NEIGHBOURHOOD RELATION

Spatial relations include many categories. To simplify our study, we take spatial neighbourhood relation as the specific subject to research its association index value.

Here we take polygon objects as example. It is usually regarded as spatially neighboured if two spatial objects share *voronoi* edge, but this definition does not give the way how to calculate the qualitative value for spatial neighbourhood relation, viz. it cannot explain Object A is more neighboured to Object O than Object B. In Section 2.2 we simply illustrate the method of influential field model to express spatial association, in the following study we use another method which can calculate spatial relation of adjacency instead of this model due to the computational complexity is heavy. Figure 3 shows Polygon 1 and Polygon 2 share common edge AB. In order to give qualitative spatial neighbourhood value, it is necessary to set a standard that can express neighbourhood index value and this value is an index for spatial association rule. Define neighbourhood index N_q for spatial objects, which do not have containing, or contained relation. N_q is positively ratio to the length of sharing edge and negatively ratio to the distance between objects canter. The central points of Polygon 1 and Polygon 2 are O1, O2, as shown in Figure2. The length of AB is l_{AB}, we will get $N_q=l_{AB}/ l_{O1O2}$, where l_{O1O2} is the distance from O1 to O2. When a study object has more than one neighbourhood objects that belong to the same type, the neighbourhood index is the sum of neighbourhood objects that sharing an edge. As show in Figure 3, in the three spatial objects Polygon 1, Polygon 2, Polygon 3, if the neighbourhood objects of Polygon 2 (they are Polygon 1 and Polygon 3) are

grouped as a same type A, then they can be merged into one. But the neighbourhood index value between Polygon 2 and type A is $\sum l/l_{O2\text{-}P}$, where O2-P is the distance from canter of Polygon 2 to the canter of its neighbourhood objects. When the index value is higher than pre-set value, then the two types of spatial objects are spatially neighboured. If two objects have containing or contained relation, they are *absolutely* neighboured. From association (here specifically neighbourhood index value), it is possible to analysed spatially associated rules both qualitatively and quantitatively. We can extend this concept to spatial buffer analysis, which is useful when we want to get the neighbourhood index for a buffer area of a given object. Although this method has its limitations, it still can be used to evaluate regular objects and some irregular spatial objects as objects having enclosure relations.

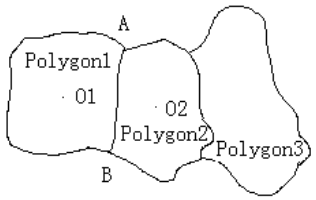

Figure 3. Association relation diagram showing adjacent neighbouring polygon objects

It is possible to get association relations other than adjacency as mentioned in Section 2.2, using the method of influential field method to calculate the association index value for separating and crossing spatial objects. Since the form of SUIT for all spatial relations are the same we will use spatial relation of adjacency (neighbourhood index value) in the following study as example.

5. SPATIAL ASSOCIATION MINING ALGORITHM

In this paper, we present a new algorithm for efficient association rule mining, which we apply in order to discover association rules in spatial databases. Our algorithm, which is called Recycled Association Rule mining (RAR), is based on the designed data structure SUIT. Other spatial relations finding are the same except a little difference of the construction of SUIT.

5.1 Description of RAR

Three steps are included in spatial-attribute comprehensive discovery: (1) to find out all large itemsets, (2) to generate rules that have confidence value higher than the predefined confidence point, and (3) to minimum rules generation. The first two steps in fact are spatial association rules mining and I/O operates on T, SR and SRV that are elements of SUIT, while step (3) is to find out comprehensive knowledge by integrating the results made from the first two steps and attribute of spatial entities and form a logically correct knowledge base or find out any logically incorrect rules from mined rules attained. In the above three steps, (1) and (3) are the key points while step (2) is relatively easy to do. Because step 3 has been introduced in our research (Bian Fuling, etc., 2001), we only give algorithm RAR for step (1).

The principle of RAR is based on a decision-tree generation method called Systematic Set Enumeration (SSE) proposed by Rymon, R (Rymon R, 1992). A decision tree is composed of tree nodes and connection line between nodes. SSE supposes all nodes have a relation of hierarchy layer structure. Except the root node, all nodes have one and only one father node. If nodes share a farther node, they are brothers. For association mining, we want to find out whether items have association relation. Every item and item combination can be regarded as node. The root node locating at the highest layer is nothing, represented as set {}. Then the node of the second layer has only one item and this layer determines the sequence of items in the decision tree. The combination of all nodes according the sequence we call sequence set (SS). Following the second layer is the third layer and every node in this layer is composed of 2 items combination. If n items are included in one transaction record, the depth of decision tree generated through SSE will be n+1 and the lowest layer has at most one node, which is composed of all items. The content of node in middle layers are made of two parts: head and tail. The head part is a item set that is completely derived from its father and the tail part is the item set of SS subtracting head item set.

The pruning strategy of decision tree can be described as following: if father node $\{N_1\}$ is not a frequent itemset then there is so use to generate its children nodes because they must not be frequent itemset. This can greatly reduce computing time.

5.2 Procedures

We use one bit segment (8 bit segment constitute one byte) to present an association flag (yes or no) and RAR to tract all possible spatial association frequent itemsets. Suppose the largest possible dimensions is *m*, the total count of record is *n*. In order to complete step 1, one time database scanning is needed with the intent to find out large itemsets and twice scanning are needed if quantitative association rules are to be mined, with the second to find out index value of association relations. The base of the implementation of RAR is on SUIT. Scanning database is meant to scan SUIT. The whole procedures is detailed in the following:

Step1: define *primary table* with two dimensions N(p1)(p2) and one-dimensioned *sum table* A(k×p3), while p1=n (total record count); p2=MOD[($\sum C_m^i$ +7)/8] (i=from 1 to m); p3=$\sum C_m^i$(i=from 1 to m). The implication of formula p2 stands for all possible nodes count of resulted decision tree with n items processed by SSE (In the equitation, /8 is to get the possible largest bytes) and p3 stands for the largest possible nodes count of resulted decision tree with n items. k is a constant value(usually 4). Every element of *sum table* records support level of corresponding elements for items in *primary table*.

Suppose the largest possible itemsets containing 10 items, then from the above step we will get the total count of bit segments is 1023 and thus p2=128, p3=1023 (note: {Ø} is not included). In the 1023 bit segments, the first 10 is the initial storage region for data import (the data stored in this region is called input attribute) and all the other segments are for temporal data storing region (the data stored in this region is called valuation attribute). For convenient purpose, the initial value of elements in both *sum table* and *primary table* are set to 0.

Step 2: Initialise the initial storage region. Fill the elements of the initial storage region by scanning SUIT. The value of elements in the initial storage region will be filled by 1 if the content of corresponding element in SUIT is not null and by 0 on the other side.

Step 3: make summary by column after the initial storage region has been initialised. The result is to fill the corresponding elements of *sum table*. If the value of the corresponding element of *sum table* is smaller than predefined support level, this column (item) is deserted because it is not

frequent itemset and will not be considered to construct higher dimension itemsets because it cannot be used to generate frequent itemsets. This step is to keep all itemsets that are impossible to form frequent itemsets out from step 4 so the whole computational complexity of RAR can be decreased.

Step 4: searching for high-dimensioned itemsets. According to the pruning strategy of SSE, two itemsets of low dimension that are frequent itemsets are select to construct a high-dimensioned itemsets. Those two selected itemsets make "and" algorithm and form support level for the high-dimensioned itemsets. If the support level is higher than the predefined one, the itemsets will be frequent itemsets. By doing like this, all possible frequent itemsets can then be found out.

Step 5: mining quantitative spatial relations. From step 1 to step 4, frequent itemsets in transaction database can be easily minded, but it is not complete for spatial rule mining because SUIT not only represent spatial association but also contains the information of index value for spatial association. In order to find out quantitatively spatially associated frequent itemsets, a second database scanning is necessary. It is regarded as quantitatively spatially associated frequent itemsets if statistical spatial index value is higher than the predefined support level.

Generally speaking, spatial associated frequent itemsets can be got by RAR. Further work will generate spatial association rules from spatially associated frequent itemsets and integrates spatial and attribute value so we can then get comprehensive knowledge.

Our emphasis here is to find out spatial association rules between spatial entities, so the whole procedures can be simplified. Spatial association relation of adjacency is entity-to-entity relation; the basic data structure that RAR is based on is a two-dimensioned table (as seen in table1). In order to illustrate RAR more clearly, we take mining spatial association relation as an example. The detailed procedures are presented in the following.

Data preparation: We translate outer data into coverage (Arc/Info data structure) because it stores the topological information, and then extract every spatial entity and its neighbouring entities and calculate the neighbourhood index value to form Spatial entities and their neighbourhood index value or SENIV (see table 2) in appliance to the neighbourhood-expressing model. Note that B and 0.5 represent the neighbouring entity name and neighbourhood index value with B respectively.

Entity	Neighbouring entity and index value
A	B:0.5; C:0.7
B	A:0.5; C:0.3; 0.4
C	A:0.7; B:0.3; A:0.4; D:0.6
D	A:0.5; C:0.6
A	B:0.4; C:0.4; D:0.5

Table 2. Spatial entities and their neighbourhood index value

Construction of SUIT: Summing neighbouring entities for each class in the same record in SENIV and their index value to form SUIT. The result of Table 1 actually comes from Table 2 after this process.

Construction of neighbourhood matrix of spatial entities: Summing each entity in SUIT according to their classification in Column T (see Table 1, entities are classified 4 types) and entity neighbourhood index value to form neighbourhood matrix of spatial entities. The result is shown in Table 3.

T	A	B	C	D
A	0	0.9	1.1	0.5
B	0.9	0	0.3	0
C	1.1	0.3	0	0.6
D	0.5	0	0.6	0

Table 3. Neighbourhood matrix of spatial entities

From table 3 we can see it is a symmetric matrix that shows the neighbourhood relations between spatial entities. The result shows that spatial class A and B, A and C have high neighbourhood index value. The value is 0.9 and 1.1 respectively.

5.3 Computational Complexity Analysis of RAR

The work is done sequentially by 5 steps. From the procedures as described in Section 5.2, we can see that the computational complexity of RAR depends mainly on Step4 and Step 5. Step1 is a constant time consuming complexity and because Step 2 is simply to scan SUIT and initialise the storage region while Step3 is simply to make summery according to the predefined classification of spatial entities so both the computational complexity of Step2 and Step 3 are $O(n)$ where n is the total number of the research entities. Step4 is to find out all frequent itemsets and its computational complexity is $O(n\log n)$. If we only consider 2-itemsets as relation of adjacency between spatial entities, the computational complexity will be $O(n)$. Step 5 scans databases again and also has to find out the frequent itemsets using the pruning strategy of SSE. So the computational complexity will be $O(n) + O(n\log n)$. The result is $O(n\log n)$. Under the worst condition, the computational complexity of RAR will be $O(n\log n)$.

6. A CASE STUDY

6.1 Background

The research areas are agricultural land in two adjacent counties in northern China. The climate of northern China is very in lack of rain and only those crops that accustomed to arid environment can survive. Those crops include peanut, cotton, maize, sorghum, etc. In reality, we make random field investigation on some crop fields by inquiring the farmers. Then we analyse the crops yields and find that some kinds of crop yields have significant difference between the two study areas. In order to find out the deep reasons that may account for the difference, we use the above proposed methods to make sure if there exists spatial association relations between planted crops. Each area of the two has an aviation image for analysis

6.2 Data Preparation

The two images are processed to extract spatial entities by the image analysis software of IDRISI. According to our study goal, we divide the planted crops into 5 categories (viz. peanut, cotton, maize, sorghum and the others). Here <the others> stands for all the other spatial entities except the mentioned 4 crops. We use supervised classification to extract the 5 kinds of spatial entities from remote sensing images. At last we convert the image data structure into vector to create crops covering polygons. We use coverage, data model used for Arc/Info, to present spatial data and use AML, the scripting language for

Arc/Info, to construct SENTV and then get SUIT. From SUIT, neighbourhood matrix of spatial entities can be built.

6.3 Generation of Spatial Association

The spatial object types have been divided into 5 and by joining the neighbourhood index value, SUIT is then built. According to the step 1 of RAR, the column of primary two-dimensioned table have 4 bytes in length and to initialise the initial storage region (5 bits). By RAR, spatial association rules can in the end be mined. The neighbourhood association rule is two-dimensioned like "A is neighbouring to B (s=70%，c=50%)". Take attribute set of spatial objects, comprehensive knowledge as if "if A is neighbouring to B, then A has higher yield (s=70%, c=50%)". For example in our study we get a rule "if cotton is planted surrounded by sorghum, it has high yield".

After extracting spatial object set that has neighbourhood association from all the objects sets, we compare the attribute (average yield) between Region A and Region B. The process is described as follows:

Suppose the spatial object sets can be divided into $C=\{C_1, C_2, \cdots, C_n\}$, and the average value of attribute A of C_i（i=1，2，…，n）in region A is X_A and X_B in region B. Let u_A, u_B and δ_A, δ_B are real value and mean square deviation of attribute A. Let $U=(X_A-X_B)/(\delta_A/m+\delta_B/n)^{1/2}$, then U is fitting to normal distribution N(0, 1). In order to verify hypothesis H: $u_A=u_B$, set confidence level *a* first. If $|U|\geqslant u_{1-0.5a}$, then abandon H, which means X_A and X_B have significant difference. In our case study, by analysing the yield of C_i(cotton) between region A and B, we found they have significant difference. And in our study, we found other factors have no significant influence on yield after the analysis of rule generation, so we can draw that it is spatial association that causes the difference. Further exploration reveals the true reason: cotton that plant surrounded by sorghum has stronger resistance ability to disease. This rule can be used to make decision on crop planting distribution in agricultural planning.

Although the spatial association rules like those mined in the case study is incidental, we can get consequential association rules from incidental production operation through comprehensive knowledge discovery. That means the valuation factor of spatial objects that are fitting to a certain rule has higher (or lower) value than those that do not show this rule. And this rule can be utilized in practice experience.

7. CONCLUSIONS

Knowledge discovered from spatial databases has been recognized as valuable knowledge acquisition in environment management, resource utilization and planning of industry and agriculture. Based on the general discussion of spatial knowledge discovery and spatial rule mining, this article proposed the principle of comprehensive knowledge discovery, concept and mining algorithm, which has a wide application in comprehensive knowledge discovering and utilization. We propose the importance to integrate mining both spatial information and their attribute description and give the way how to attain this goal through theoretical analysis and case study. Although the comprehensive knowledge discovery proposed here focuses on spatial association rule mining and attribute data, it can also be applied in other comprehensive knowledge discovery areas such as spatial classification, spatial clustering, etc., which will be included in our next research.

REFERENCES

Agrawal R, Imielinski T, Swami A, 1993. Mining association rules between sets of items in large databases. In: *Proc. of the 1993 ACM-SIGMOD: International Conference on management of Data*, Washington, DC., pp. 207~216.

Bian Fuling, Sha Zongyao, Chen Jiangping, 2001. Generation of Minimum Rules from Rough Rule Sets Based on Object-Spatial Information Table. *Journal of Wuhan Uninversity (information version in Chinese)*, 26(5), pp. 399~403.

KERRY TAYLOR, GAVIN WALKER and DAVE ABEL, 1999. A framework for model integration in spatial decision support systems. *Int. J. Geographical information science*, 13(4), pp. 533~555.

Leung, Yee, Ma Jiang-Hong, Zhang Wen-Xiu, 2001. New method for mining regression classes in large data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1), pp. 5-21.

Rymon R, 1992. Search through systematic set enumeration. In: *Proceeding of the 3rd International Conference on Principle of Knowledge Representation and Reasoning*. Massachusetts, USA. pp. 539~550

Sauchyn, DJ, Yong Xongchao, 1991. Structure and application of URIFT: An interface between image analysis and vector GIS. *Canadian Journal of Remote*, 17(4), pp. 332-338.

S. Lavingto N, Dewhurst E, Wilkins A Freitas, 1999. Interfacing knowledge discovery algorithms to large database management systems. *Information and software technology*, 41(9), pp. 605~617

S Rahayana, A Siberschatz, 1998. On the discovery of interesting patterns in association rules. In: *Proc. of the 24th VLDB Conference*, New York, USA, pp. 368~379

T.Q. Zeng and Q. Zhou, 2001. Optional spatial decision making using GIS: a prototype of a real estate geographical information science. *Int. J. Geographical information science*, 2001, 15(4), pp. 287~306.

W. Lu, J. Han and B. C. Oci, 1993. Discovery of General Knowledge in Large Spatial Databases. In: *Proc. Far East Workshop on Geographic Information Systems*, Singapore, pp. 275~289.

Ziarko W, 1995. Introduction to the special issue on Rough sets and knowledge discovery. *International Journal of Computational Intelligence*, 11(2), pp. 223~226.

Zhang T, Ramakrishnan R, Livny, 1997. BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery*, 3(1), pp. 141~182.

ACKNOWLEDGEMENTS

This study was partially supported by the China's National Surveying Technical Fund (No. 20007). The authors also deeply appreciate the constructive suggestions made by the paper reviewers to polish this article.