Method and Application of Comprehensive Knowledge Discovery

SHA Zongyao  BIAN Fuling

1 Introduction

Knowledge discovery from databases (KDD) or data mining (DM) has been an area of increasing interests in recent years, esp. in the field of machine learning, pattern recognition, statistics, artificial intelligence, and high performance computing[1-3]. The development of GIS has eased the management of spatial data, which carries topological and/or direction and/or distance information and which is usually organized in sophisticated, multi-dimensional spatial indexing structures. But knowledge discovery from spatial databases imposes a difficult task on data managers. The discovered knowledge can be grouped into spatial information generalization, spatial association, spatial classification and spatial clustering. Both experience and theory study prove that knowledge discovery from spatial databases (KSDS) can not be independent of spatial objects attributes and spatial objects and their attributes should be connected together to give a complete expression of spatial objects. It is very important to construct a suitable spatial data-mining model that can ease the process of KSDS.

2 Theory and concepts

2.1 Spatial knowledge expression system

Let $S = (U, C, D, V, f)$ and $U = \{u_1, u_2, \ldots, u_n\}$, where $U$ is a finite set of objects; $A = C \cup D$ is the attribute set; $C = \{a_1, a_2, \ldots, a_m\}$ is the condition attribute set.
The attention should be taken that C contains spatial constraint conditions; \( D = \{d_1, d_2, \ldots, d_s\} \) is the decision attribute set; \( V \) is the field domain set composed of \( C \cup U \), viz. \( V = \bigcup_{p \in A} V_p \), \( V_p \) is the value domain of field \( p \); \( f \) is an information function, viz. \( f : U \times A \rightarrow V \). And \( S \) is defined as the formalization definition of the spatial knowledge expression system (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 the 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 the 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 \( n \)-dimensional set will be given spatial index value (fuzzy index type) or 0-1 (Boolean index type) according to the spatial association.

2.2 Some concepts

2.2.1 Comprehensive knowledge discovery

Comprehensive knowledge discovery is to analyse comprehensively the spatial characters as well as the attributes of spatial entities and to find out the deep regulations that are stored implicitly in the attribute information and spatial information of their objects. For example, in the process of analysing the influential factors upon the crop yield, we should not only consider the possible attribute factors such as climate, soil fertility, soil texture, etc., but also consider the spatial information (climate and soil fertility distribution) that may contain unknown spatial association patterns which can then be used to support the decision for crop planting area planning and yield evaluation.

2.2.2 Spatial union information table (SUIT)

SUIT is defined as the 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 the classification and relations of spatial entities and attribute information (AI) which records the attribute fields of spatial entities. In a formalized form, SUIT is expressed as SUIT \( (T, SR, SRV, AI, AIV) \), here \( T \) stands for the whole set of the spatial entities, and \( SRV \) and \( AIV \) are the index value or the representative mode of spatial relations and the attribute value of the spatial entities. For specific purposes, some simplifications have to be made. Let \( SUIT' \) be the actual study goal, and \( T' \), \( SR' \), and \( AI' \) be the subsets of spatial objects, the spatial relation objects and the attribute of the spatial objects respectively, viz. \( T' \subseteq T \), \( SR' \subseteq SR \), \( AI' \subseteq AI \). As shown in Figure 1, the sub-sets of the spatial objects are; \( T' = \{A, B, C, D, A\} \), the classification of spatial objects is; \( SR' = \{A, B, C, D\} \), suppose the attribute field sets \( AI' = \{\text{area, perimeter}\} \), 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 Fig. 1. \( SUIT' \) will be used to stand for SUIT in the following sections.

![Spatial object relation](image)

Table 1 SUIT' for spatial objects

|     | 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  |
We suggest a method called influential field for qualitatively representing the spatial relations by extending the adjacency relation. In this method, the spatial objects need not share edges (in Fig. 1, the spatial objects share edges). They can be crossing, separated or adjacent. Fig. 2 shows two separated spatial objects $O_1$ and $O_2$, and their influential fields. Numerous fields caused by all objects influence each object in a given area. Those spatial objects can also be classified as A, B, etc., thus for any object, its relation index value with other objects can be calculated by the influential field model. In this way, the possible relations between the spatial objects can be extended.

![Influential field for two separated spatial objects](image)

3 Overview of association rule and spatial association rule mining

3.1 Association rule and spatial association rule

Association rule describes the item relations in a database. In a mathematic language, let $I = \{i_1, i_2, \ldots, i_n\}$, which is an itemset called dataset, and let $D$ be a collection of all possible itemsets.

Transaction $T$ is a subset of $I$, viz. $T \subseteq I$, and every transaction is identified by TID. For dataset $X$, we call that $T$ includes $X$ if and only if $X \subseteq T$. The association rule is usually presented in the shape like; $X \Rightarrow Y$, here $X \subseteq I, Y \subseteq I$, and $X \cap Y = \emptyset$. Here, $X$ is called the condition of the rule while $Y$ is the result from the condition $X$. The confidence level for the rule $X \Rightarrow Y$ on the basis of set $D$ is defined as $\alpha \%$ which means that, out of all transactions in $D$, $\alpha \%$ records include both $X$ and $Y$, and the support level of rule $X \Rightarrow Y$ is defined as $\beta \%$ that means $\beta \%$ out of all transaction records include $X \cup Y$. Rules whose support level is higher than the predefined support level are called the frequent rules, and both confidence level and support level, higher than the predefined ones, are called the intensive rules.

The focus of spatial association rule mining is on spatial information. In the 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, we get $A(R)B$ if $A$ and $B$ have a spatial association relation $R$. For example, if 80% of all area in the width of five kilometers along the sides of river $A$ is distributed by agricultural fields then we say the agricultural fields and the river have a certain association relation.

3.2 Review of algorithms in mining association rules

The general method for mining the association rules includes the following procedures: 1) to find out all frequent itemsets; 2) to form the association rules from the frequent itemsets. For the given full item set $U$, if $A \subseteq U$ is a subset of $U$ and $sup(A)/sup(U) > Confidence$, where $sup(X)$, $Confidence$ stand for the support level and the confidence level for the spatial type $X$, then the rule $A \Rightarrow U-A$ can be induced. In the above two steps, the first step is key. Once the frequent itemsets have been obtained, 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 the transaction databases. Because Apriori is very time consuming in generating
large itemsets, a hashing pruning method called DHP algorithm was proposed.

Besides the above-mentioned algorithms, the generalized association rule, the multi-level association rules, and the quantitative association rules mining were proposed. But nearly all of those algorithms need scanning database so many times that their efficiency is greatly reduced.

All of those methods are directed to the transaction database, however. As for the spatial association rule mining, those methods can also be applied after a little modification, but the spatial database must be based on suitable data model and even so, the efficiency may be low. In Section 5, we will give an efficient algorithm (RAR) to the find spatial association rules.

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 gain the association index value. Here, we take polygon objects as an example. It is usually regarded as spatially neighbouring 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 that Object A is more neighbouring to Object O than Object B. Fig. 3 shows that Polygon 1 and Polygon 2 share a common edge AB. In order to give the qualitative spatial neighbourough value, it is necessary to set a standard which can express the neighbourhood index value for the spatial association rule. Defining neighbourhood index \( N_q \) for spatial objects, which do not have containing or contained relation, \( N_q \) has positive correlation to the length of sharing edge and has negative correlation to the distance between objects’ centres. The central points of Polygon 1 and Polygon 2 are \( O_1 \), \( O_2 \), respectively (Fig. 3). The length of AB is \( l_{AB} \), then we can obtain \( N_q = l_{AB}/l_{O_1O_2} \), where \( l_{O_1O_2} \) is the distance from \( O_1 \) to \( O_2 \). When a study object has more than one neighbourhood object belonging to the same type, the neighbourhood index is the sum of neighbourhood objects sharing an edge. As shown in Fig. 3, in the three spatial objects, Polygon 1, Polygon 2, Polygon 3, if the neighbourhood objects of Polygon 2 (i.e. Polygon 1 and Polygon 3) are grouped as the same type \( A \), then they can be merged into one. However, the neighbourhood index value between Polygon 2 and type \( A \) is \( \sum l/l_{O_2P} \), where \( O_2-P \) is the total distance from centre of Polygon 2 to the centre of its neighbourhood objects. When the index value is higher than the pre-set value, then the two types of spatial objects are spatially neighbouring. If the two objects have containing or contained relation, they are absolutely neighbouring.

5 Spatial association mining algorithm

In this section, we present a new algorithm for efficient association rule mining, which we apply 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.

5.1 Description of RAR

The spatial-attribute comprehensive discovery includes three steps: (1) to find out all large itemsets, (2) to generate rules that have a confidence value higher than the predefined
confidence value. 2 to minimize the rule 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 from the first two steps and the attribute of the 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, Steps 1 and 3 are key while Step 2 is relatively easy to do. Step 3 has been introduced in our previous research (4), so we only give algorithm RAR for Step 1.

5.2 Procedures

We use one bit segment (eight bit segments constitute one byte) to represent an association flag (yes or no) and RAR to calculate all possible spatial association frequent itemsets. Supposing that the largest possible dimension is $m$, the total number of record is $n$. In order to complete Step 1, once database scanning is needed to find out large itemsets and twice scanning is needed if quantitative association rules are to be mined. The second scanning is to find out the index value of the association relations. The base of the implementation of RAR is on SUIT. Scanning database means scanning SUIT. The whole procedures are detailed as follows:

Step 1: Define primary table with two dimensional $N(p_1)(p_2)$ and one-dimensional sum table $A(k \times p_3)$, while $p_1 = n$ (the total number of record); $p_2 = \text{MOD}[(\sum C^i_n + 7)/8]$ ($i = 1, 2, \cdots, m$); $p_3 = \sum C^i_n$. Equation $p_2$ stands for all possible nodes count of resultant decision tree with $n$ items processed by Systematic Set Enumeration (SSE) proposed by Rymon R (5) and $p_3$ stands for the largest possible nodes count of resultant decision tree with $n$ items and $k$ is a constant value (usually four). Every element of sum table records support the level of the corresponding element for the items in primary table.

Supposing that the largest possible itemset contains 10 items, then following the above step we will obtain the total count of bit segments which is 1 023 and thus $p_2 = 128$, $p_3 = 1 023$ (note: $\{\phi\}$ is not included). In the 1 023 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 the temporal data storing region (the data stored in this region is called valuation attribute). For convenience, the initial values of elements in both sum table and primary table are set to 0.

Step 2: Initialize 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 one if the content of corresponding element in SUIT is not null and will be filled by zero on the other side.

Step 3: Make summary by column after the initial storage region has been initialized. The result is filled to the corresponding elements of sum table. If the value of the corresponding element of sum table is smaller than the predefined support level, this column (item) is deserted because it is not frequent itemset, and will not be considered in the construction of the higher dimension itemsets because it cannot be used to generate frequent itemsets. This step is to keep all itemsets which are impossible to form frequent itemsets out of consideration in the next step so the whole computational complexity of RAR can be decreased.

Step 4: Search for high-dimensional itemsets. According to the pruning strategy of SSE, two itemsets of low dimension that are frequent itemsets are selected to construct a high-dimensional itemsets. Those two selected itemsets make A algorithm and form the support level for the high-dimensional itemsets. If the support level is higher than the predefined one, the itemsets will be frequent itemsets. In this way, all possible frequent itemsets can then be found out.
Step 5: Mine quantitative spatial relations. From Step 1 to Step 4, the frequent itemsets in transaction database can be easily obtained, but it is not complete for spatial rule mining because SUIT not only represents the spatial association but also contains the information of index value for the 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 the statistical spatial index value is higher than the predefined support level. Our emphasis is on finding out spatial association rules between spatial entities, so the whole procedures can be simplified. The spatial association relation of adjacency is the entity-to-entity relation; the basic data structure that RAR is based on is a two-dimensional table (Table 1).

To illustrate RAR more clearly, we take the spatial association mining as an example. The detailed procedures are presented as follows.

(1) Data preparation: We translate the outer data into coverage (Arc/Info data structure) because in it the topological information is stored, 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) according to the neighbourhood-expressing model. Note that B and 0.5 represent the neighbouring entity name and the 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, A:0.4               |
| C      | A:0.7, B:0.3, A:0.4, D:0.6        |
| D      | A:0.5, C:0.6                      |
|        | B:0.4, C:0.4, D:0.5               |

(2) Construction of SUIT: Summing the 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.

(3) Construction of the neighbourhood matrix of the spatial entities, Summing each entity in SUIT according to their classification in Column T (see Table 1, entities are classified into four types) and the entity neighbourhood index value to form the neighbourhood matrix of the 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  |

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

5.3 Computational complexity analysis of RAR

From the procedures described in Section 5.2, we can see that the computational complexity of RAR depends mainly on Step 4 and Step 5. Step 1 is of a constant time consuming complexity. Because Step 2 is simply to scan SUIT and initialise the storage region while Step 3 is simply to make summary according to the predefined classification of the spatial entities, both the computational complexity of Step 2 and Step 3 are O(n) where n is the total number of the research entities. Step 4 is to find out all frequent itemsets and its computational complexity is O(nlogn). If we only consider two itemsets as the relation of adjacency between the spatial entities, the computational complexity will be O(n). Step 5 the 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(nlogn). The result is O(nlogn). 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 lands in two adjacent counties in north China. North China is very short of rain and only those crops that are accustomed to arid environment can survive. In reality, we make random field investigation in some crop fields by inquiring the farmers, and then we analyse the crops yields and find that some kinds of crop yields are of significant difference between the two study areas. In order to find out the reasons that may account for the difference, we use the above proposed methods to make sure if there exist the spatial association relations between the planted crops. Each area of the two has an aerial image for analysis.

6.2 Data preparation

The two images are processed to extract the spatial entities by the image analysis software of iDRISI. According to our goal, we divide the planted crops into five categories (viz. peanut, cotton, maize, sorghum and the others. Here, the others refer to all the other spatial entities except the mentioned four crops). We use the supervised classification to extract the five kinds of spatial entities from remote sensing images. Last, we convert the image data structure into vector to create crop covering polygons. We use coverage data model to present spatial data and AML, the scripting language for Arc/Info, to construct SUIT. From SUIT, the neighbourhood matrix of the spatial entities can be built.

6.3 Generation of spatial association

The spatial object types have been divided into five categories and by adding the neighbourhood index value, SUIT is then built. According to the Step 1 of RAR, the column of primary two-dimensional table have four bytes in length and then the initial storage region (five bits) is initialized. By RAR, the spatial association rules can be obtained. The neighbourhood association rule is two-dimensional like "A is neighbouring to B ($s = 70\%, c = 50\%)". By taking the attribute set of the spatial objects, the comprehensive knowledge can be explained as "if A is neighbouring to B, then A has higher yield ($s = 70\%, c = 50\%)". In our study, for example, we obtain a rule that "if cotton is planted surrounded by sorghum, it has higher yield".

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

Supposing that the spatial object sets can be divided into $C = \{C_1, C_2, \cdots, C_n\}$, and the average value of attribute $X$ of $C_i$ ($i = 1, 2, \cdots, n$) in Region A is $X_A$ and $X_B$ in Region B. Let $u_A, u_B$ and $\delta_A, \delta_B$ are real values and mean square deviations of attribute $X$. Let $U = (X_A - X_B)/\sqrt{(\delta_A/n + \delta_B/n)}$, then $U$ is fitting to the normal distribution $N(0, 1)$. In order to verify the hypothesis $H: u_A = u_B$, the confidence level $a$ is set firstly. If $|U| \geq u_{1-a/2}$, then $u_A = u_B$ is abandoned, which means $X_A$ and $X_B$ have significant difference. In our case study, by analysing the yield of $C_i$ (cotton) between Regions A and B, we found that they have significant difference and other factors have no significant influence on yield after the analysis of rule generation. So we can draw that it is the spatial association that causes the difference. The further exploration reveals the true reason, cotton planted around sorghum has stronger resistance ability to disease. This rule can be used to make decision in the crop planting distribution in agricultural planning.

7 Conclusions

The knowledge discovered from the spatial
databases has been recognized as the valuable knowledge acquisition in the environment management, the resource utilization and the planning of industry and agriculture. On the basis of the general discussion of the spatial knowledge discovery and the spatial rule mining, this paper gives the principle of the comprehensive knowledge discovery, concept and mining algorithm, which has a wide application in comprehensive knowledge discovering and utilization. It is important to integrate mining both the spatial information and the attribute information. The theoretical analysis and the case study are given to attain this goal. Some points are summarized as below.

1) Data mining should consider both spatial relation and attribute character of spatial objects, which is called comprehensive knowledge discovery.

2) Spatial relation can be described from different views and only some key factors that have important influence on our study domain can be considered.

3) Spatial association relation and attribute character of spatial objects are researched in the case study and valuable patterns are obtained.

4) A comprehensive data-mining algorithm is introduced and used in this paper.

Although the comprehensive knowledge discovery proposed here focuses on the spatial association rule mining and attribute data, it can also be applied to other comprehensive knowledge discovery fields such as spatial classification, spatial clustering, etc., which will be included in future researches.

References

1 Yee L, Ma J H, Zhang W X (2001) New method for mining regression classes in large data sets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(1): 5-21

2 Lu W, Han J, Oci B C, (1993) Discovery of general knowledge in large spatial databases. Far East Workshop on Geographic Information Systems, Singapore.

3 Taylor K, Walker G, Abel D (1999) A framework for model integration in spatial decision support systems. Int. J. Geographical Information Science, 13 (4): 533-555

4 Bian F L, Sha Z Y, Chen J P (2001) Generation of minimum rules from rough rule sets based on object-spatial information table. Geomatics and Information Science of Wuhan University, 26(5): 399-403

5 Rymon R (1992) Search through systematic set enumeration. The 3rd international Conference on Principle of Knowledge Representation and Reasoning, Massachusetts, USA.

(Continue from Page 19)

5 Ban Y F (1995) Improving the accuracy of synthetic aperture radar analysis for agricultural crop classification. Canadian Journal of Remote Sensing, 21 (2), 158-164

6 Barber D G, Drew E F (1991) SAR sea ice discrimination using texture of statics, a multivariate approach. Photogrammetric Engineering and Remote Sensing, 57(4), 385-395

7 Ulaby F T, Kouyate F (1986) Textural information in SAR image. IEEE Transactions on Geoscience and Remote Sensing, 24(2), 234-245