STUDY ON SCALE TRANSFORMATION IN SPATIAL DATA MINING

Qingxian Sun[1,2], Tao Fang[2], Dazhi Guo[1]

[1]Department of surveying and land science, China University of Mining & Technology, Beijing, China, 100083, 6886sun@163.com Tel: 86-010-51733057/86-021-62933787
[2]Institute of Image Processing & Pattern Recognition, Shanghai Jiao Tong University, No.1954, Huashan Road,Shanghai,China,200030, tfang@sjtu.edu.cn,Tel: 86-021-64472192

KEY WORDS: scale, SDMKD, spatial association rules mining, spatial database, geostatistics

ABSTRACT:

It is well known that a great deal of achievement about data mining in business field has been made during the past years. In this case, the research emphasis on data mining and knowledge discovery has inevitably been shifted from non-spatial data into spatial data. A marked and important difference compared non-spatial data with spatial data is of scale-dependency for spatial data, i.e., its scale attribute and nature. In this paper, we initially present the scale issue of spatial data mining, and conduct an experiment to briefly demonstrate scale transformation. The results show that information and knowledge may be changeful with the scale changing. Simultaneously, to solve our problem, we successfully introduce geostatistics into SDMKD.

1. INTRODUCTION

Nowadays, large amount of spatial data have been collected from many applications and data collection tools. “The spatial data explode but knowledge is poor”[Li et al.,2002], therefore, “We are drowning in data, but starving for knowledge!” The implicit knowledge hidden in those spatial data cannot be extracted using traditional database management systems. To solve this problem, a sub-field of knowledge discovery in large database, called spatial data mining and knowledge discovery(SDMKD), has been introduced. The objective of SDMKD is to discover interesting patterns in large spatial databases which is more and more interested by many researchers.

The early research of data mining concentrated on non-spatial data, after great achievement in business, insurance, finance, etc., research emphasis has been inevitably shifted from non-spatial data(market basket data) to spatial data.

We all know that there is marked and important difference between non-spatial data spatial data: all spatial datum are governed by scales in nature. To characterize a spatial object precisely and accurately, it is essential to state exactly the concomitant scale of the spatial objects.

Scientific literatures in related fields of earth sciences clearly indicate an increasing emphasis on studies at multiple scales. For example, multi-scale is a hot topics in GIS, many scholars are concentrating on the focus. Thus, translating information from one scale to another, i.e., scale transformation(scaling), is indispensable. In fact, in the fields of ecology, remote sensing, environmental science and so on, scale transformation is serving as a process technique to deal with datasets and becoming a forefront research focus.

Though scale is crucial to all study fields related to earth sciences, and many scholars have been realized its importance, few researchers in the field of SDMKD have done works up to now.

In this paper, we initially present the scale issue of spatial data mining, and sure that it will become a new important direction. Firstly, the concept of "scale" and "scale transformation" will be described. Secondly, we introduce algorithm for our experiment and data processing in detail. Finally, we explain the results and draw a conclusion.

2. SCALE TRANSFORMATION IN SDMKD

2.1 About the term “scale”

The term scale is a mistakable word for its changeability in meaning. In the field of geodesy, cartology and geography, scale is defined as a ratio between distances measured on the map and on the ground. In the field of math, electronics, optics and mechanics, scale is defined as measuring tool or filter. In the field of ecology, coarse scale indicates broad area and long time period and fine scale indicates the quite reverse. The notion of scale in remote sensing means automatically the resolution (cell size).In short, scale is a window of perception though which we observe the investigated objects[Guo Dazhi et al.,2003].

On the one hand, there are two facets to scale, namely spatial scale and temporal scale. On the other hand, to express scale exactly, two methods are often used, they are grain and extent[Wu J.,1999].

2.2 Scale transformation in SDMKD

In fact, the scale issue is a crucial problem in the earth science and related fields. Many researchers have greatly devoted endeavor and passion to this domain. It is understood that some information and knowledge mined from spatial data may be changeful due to showing different shape, structure and detail under different scales. Some knowledge (rules or patterns) could be constant, some changeable, and perhaps, some disappear whereas some new knowledge appear unpredictably. Spatial rules without exact scale are invalid, even result in loss in some cases.

Deducing information from macro-scale to small-scale is called up-scaling, deducing from small-scale to macro-scale down-scaling, and deducing between same scales iso-scaling. In general, enlarging the scale can increase the quantity of information hidden in spatial datasets, whereas shrinking the

scale can decrease the quantity of information, but notice that this is non-linear. In the field of SDMKD, changing scale does not only mean increase or decrease of information, but also means shift and disappear of information. Clearly, study on how to transfer information among different scales is one of the necessary task in SDMKD, actually which is also urgent.

3. DESIGN APPROACH OF A STUDY CASE

The above idea is only our guess theoretically. To testify the guess, an experiment is conducted. A piece of suburban farm field alongside of a sewage canal is polluted by several kinds of heavy metals due to long time irrigation using wastewater. For the purpose of soil environmental quality assessment, soil-chemists sampled topsoil from the polluted field in terms of definite rules specified in advance, encoding as well. These samples were taken back to laboratories to be tested. The contents of heavy metals of each sample were thus verified and then stored in computer. We just use the results to achieve our goal. Figure1 is the graph of the experiment field. Farm area is about 0.2km^2. Interval of sample is 30m. Total number of samples is 199.

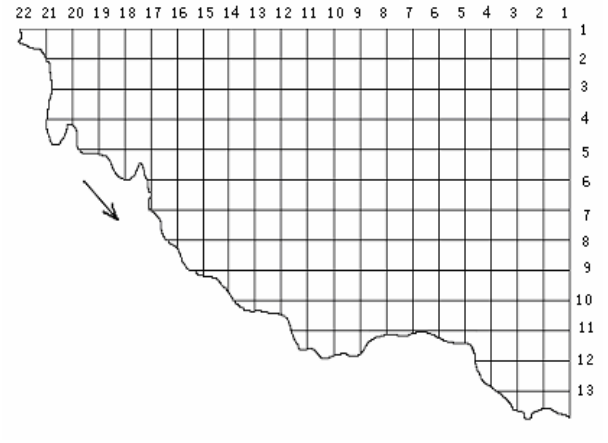


Figure1. sampling plot (intersection for sampling)

Our technical design is to make use of these records to generate association rules among contents of heavy metals. The technical method is to take a record of every other sample of two directions, i.e., double the interval, therefore sample density is half of original, thus, under the circumstances there are 56 records. Using the 56 records, we form a database. Consequently we form another database using the 199 original records. In our experiment, grain(sample density) means scale, thus changing grain is also equal to changing scale. The first database scale is just half of the second one. Below is our design approach in detail:

(1) First of all, data pre-processing. In this paper, we use Apriori algorithm, which deals only with "market basket data"(Boolean), to mine our datasets. For quantitative attributes, we need to pre-process the data to fit the Boolean association rule mining model. A simple way is to partition the values of the quantitative attribute into intervals and then combine adjacent intervals as necessary. This procedure is also called discretization.

(2) Then, transforming pre-processed records into transactional databases by two scales, namely, two transactional databases with different grain respectively. And

(3) Finally, mining knowledge and analyzing results.

4. ASSOCIATION RULE AND ITS ALGORITHM

There are many data mining techniques, such as association rule mining(ARM), classification, clustering, sequential pattern mining, etc. Association rule mining, first introduced in 1993 by R. Agrawal et al., is an important technique, and it has been extensively studied and applied for data analysis in many fields. The task of association rule mining is to find certain association relationships among a set of data items in a database. The association relationships are described in association rules. Originally, association rule mining focused on "market basket data" which store items purchased on a per-transaction basis. A typical example of an association rule on "market basket data" is that 70% of customers who purchase bread also purchase butter. At present, spatial association rule mining is becoming a prospective focus. An example of spatial association rule might be "most big cities in Canada are close to the Canada-U.S. border."

A formal statement of the association rule problem is [Agrawal1993,1994] [Cheung1996]:

Definition 1: Let I ={I_1, I_2, … , I_m} be a set of m distinct attributes, also called literals. Let D be a database, where each record (tuple) T has a unique identifier, and contains a set of items such that T⊆I. An association rule is an implication of the form X⇒Y, where X, Y⊂I, are sets of items called itemsets, and X∩ Y=ϕ. Here, X is called antecedent, and Y consequent. There are two important measures, support (s) and confidence (c), used in assessing the quality of the rules. They can be defined below.

Definition 2: The support (s) of an association rule is the probability (often in percent) of the records that contain X∪ Y to the total number of records in the database.

$$s(A \Rightarrow B) = \text{support}(A \cup B)$$

Therefore, if we say that the support of a rule is 5% then it means that 5% of the total records contain X∪ Y. Support is the statistical significance of an association rule.

Definition 3: For a given number of records, confidence (c) is the conditional probability (often in percent) of the number of records that contain X∪ Y to the number of records that contain X.

$$c(A \Rightarrow B) = \text{support}(A \cup B)/\text{support}(A)$$

Thus, if we say that a rule has a confidence of 85%, it means that 85% of the records containing X also contain Y. The confidence of a rule indicates the degree of correlation in the dataset between X and Y. Confidence is a measure of a rule's strength. Usually a large confidence is required for association rules.

The goal of association rule mining is to find all the rules with support and confidence exceeding user specified thresholds. There several algorithms to find association rules, such as Apriori, DHP(direct hashing and pruning), DIC(dynamic itemset counting), FP-growth(frequent pattern), and Partition, among which Apriori is the most important one[Qin Ding, 2002]. The Apriori algorithm developed by R.Agrawal in 1994 is a great achievement in the history of mining association rules[Cheung1996]. It is by far the most well-known association

rule algorithm. In this paper, we use the famous Apriori algorithm to perform our design.

5. DATA ORGANIZATION AND DISCRETIZATION

In our experiment, we select Cu, Cr, Pb and Zn as the subjects. The original form is relational database given in Table 1(only 3 samples are selected randomly to show). Accordingly, each sample is a transaction while each metal is an attribute.

Sample code	Cu	Cr	Pb	Zn
1	111.3	94.5	24.3	145.7
2	135.6	114.7	62.1	307.2
3	38.3	74.1	40.9	90.8

Table 1 Test results of heavy metal element content in polluted topsoil(mg/kg)

Apriori algorithm is a Boolean model while datum in Table 1 are consecutively quantitative. To solve the problem, we can use the data discretization technique. Data discretization, also called data partitioning, is to discretize the values of consecutive attributes into a smaller number of intervals, where each interval is mapped to a discrete symbol. According to assessment standard of heavy metal to soil pollution(Table 2), we use user-defined partitioning to discretize our datum. In this case, we contrast each sample data one by one with pollution degree to classify its category. Take the first sample for example, 4 kinds of contents belong to Ⅰ, Ⅰ,None(non-polluted) and Ⅰ. Hereby Table 1 can be converted into another form showed in Table 3.

Pollution degree	Cu	Cr	Pb	Zn
Ⅰ(slight-polluted)	28.6	75.0	25.1	93.0
Ⅱ(moderate-polluted)	125.0	300.0	200.0	300.0
Ⅲ(severe-polluted)	400.0	400.0	500.0	500.0

Table 2 Assessment standard of heavy metal in topsoil(mg/kg)

Sample code	Cu	Cr	Pb	Zn
1	Ⅰ	Ⅰ	None	Ⅰ
2	Ⅱ	Ⅰ	Ⅰ	Ⅱ
3	Ⅰ	None	Ⅰ	None

Table 3 Results of discretization by user-defined partitioning

Even so, Table 3 is not compatible with Apriori model. We have to get on with more transformation. Let result of 4 pollution degree of Cu correspond to consecutive integers, i.e., None to 0, Ⅰ to 1, Ⅱ to 2, and Ⅲ to 3, respectively. Subsequently, let result of 4 pollution degree of Cr correspond to the next consecutive integers, viz., None to 4, Ⅰ t 5, Ⅱ to 6, and Ⅲ to 7, respectively. So do the next two. Ultimately, each data in Table 1 is mapped to integer. Thus the final database showed in Table 4, can be used by the Apriori model directly. The whole procedure of transformation can be finished automatically by program.

Sample code	Cu	Cr	Pb	Zn
1	1	5	8	13
2	2	5	9	14
3	1	4	9	12

Table 4 Final result of data discretization

The first database is of 60m grain(sample interval) while the second one is of 30m grain(sample interval) in scale.

6. EXPERIMENT RESULTS

6.1 Mining results from two databasesets

We test our approach with the two databases. The association rules are discovered setting a minimum support at 25% and the minimum confidence at 80%. Rules mined from two databases are listed in Figure2.

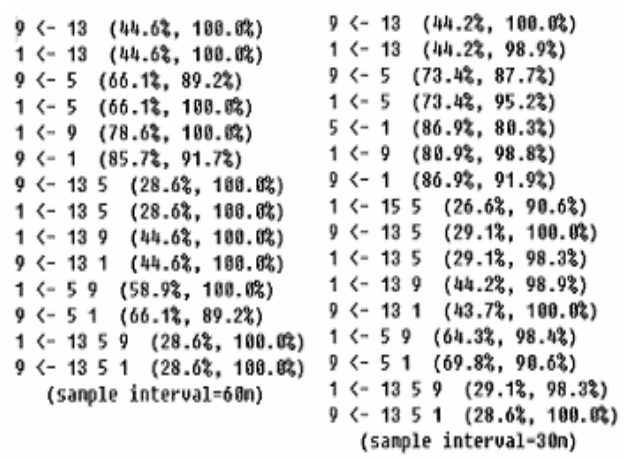


Figure2.Rrules mined from two different density databases

Experiment results validate our idea. From first database (interval=60m) and second database(interval=30m), we mine 14 and 16 rules respectively. Most rules remain or change with slight fluctuant support and confidence, such as 9 <= 13 5 1 (28.6%, 100.0%). Two new rules appear, they are 5 <= 1 (86.9%, 80.3%)and 1 <= 15 5 (26.6%, 90.6%). No rules disappear during this course.

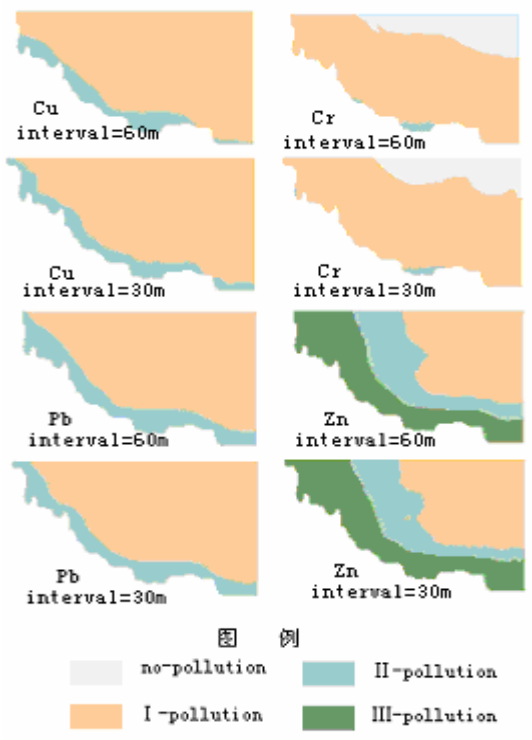
Rules can be interpreted as following:
Two-member rule, for example, 13 $\Rightarrow$ 9(44.6%, 100%), 44.6% samples are polluted by Zn and Pb simultaneously with Ⅰ degree, and if the sample is polluted by Pb(Ⅰ), the probability of this sample that polluted by Zn(Ⅰ) is 100%.

Three-member rule, for example, 5 ∧ 3=>9(29.1%,100%), 29.1% samples are polluted by Cu(Ⅲ), Cr(Ⅰ) and Pb(Ⅰ) simultaneously, and if the sample is polluted by Cu(Ⅲ) and Cr(Ⅰ) simultaneously, the probability of this sample that is polluted by Pb(Ⅰ) is 100%.

Three-member rule, for example, 1 ∧ 5 ∧ 13=>9(28.6%,100%),there are 28.6% samples are polluted by Zn(Ⅰ), Cr(Ⅰ), Cu(Ⅰ) and Pb(Ⅰ) simultaneously, and if the sample is polluted by Zn(Ⅰ), Cu(Ⅰ) and Cr(Ⅰ) simultaneously, the probability of this sample is polluted by Pb(Ⅰ) is 100%.

For the purpose of reasonable explanation for the results, we generate a pollution degree distribution graph(Figure3) by Kriging interpolation through using GIS. We all know that

Kriging interpolation is the core part in geostatistics, and this method is particularly suitable for mineral resources prediction, geological exploration survey, soil component analysis. From the graph we can get more information for the results, the key course of the method is semivariance analysis.



Figuer3.Pollution degree distribution graph by Kriging interpolation

6.2 Predictions of association rules and pollution degree graph

According to trend of pollution degree distributions, we imitate a set of semivariograms by spheric model for metal of Cu,Cr,Pb and Zn respectively, their parameters are listed in Table 5. In the light of these semivariances, we predict pollution degree distribution under 15m sample interval(Figure4), association rules((Figure5) as well. There are 726 hypothetic samples under 15m interval.

	Cu	Cr	Pb	Zn
Sill(C_1+C_0)	0.49	0.38	0.45	0.40
Nugget(C_0)	0.001	0.015	0.033	0
Range(m)	685	658	637	637

Table 5. Parameters of spheric model of the theoretic semivariogram

Figure4. Pollution degree distribution prediction under 15m sample interval(legend is same as Figure3)

```
5 <- 15   (25.3%, 81.0%)
1 <- 15   (25.3%, 88.0%)
9 <- 13   (49.9%, 100.0%)
1 <- 13   (49.9%, 100.0%)
9 <- 5   (73.1%, 94.4%)
1 <- 5   (73.1%, 99.2%)
1 <- 9   (91.9%, 100.0%)
9 <- 1   (97.0%, 94.7%)
9 <- 13 5   (29.2%, 100.0%)
1 <- 13 5   (29.2%, 100.0%)
1 <- 13 9   (49.9%, 100.0%)
9 <- 13 1   (49.9%, 100.0%)
1 <- 5 9   (69.0%, 100.0%)
9 <- 5 1   (72.6%, 95.1%)
1 <- 13 5 9   (29.2%, 100.0%)
9 <- 13 5 1   (29.2%, 100.0%)
```

Figure5. Prediction association rules under 15m sample interval

Comparing rules under 15m interval and 30m interval, we find quantity of rules remains but content of rules changes slightly. Two new rules appear, and two rules disappear at the some time.

7. SUMMARY AND CONCLUSION

We have presented issue of scale in SDMKD and developed a method for mining spatial association rules. Experimental results demonstrate that spatial association rules often change with different scales, i.e., spatial association rule is of scale dependency. The results also indicate that rules, whose support and/or confidence are close to minimum support and/or minimum confidence, are more sensitive to scales. Conventional theory is powerless to the problem, and thus we use geostatistics in the course of experimenting to form database. In fact, issue of scale in SDMKD is more than what we have done, the present research is only a good start.

ACKNOWLEDGEMENTS

The work described in this paper is supported by the funds from National Natural Science Foundation of China (Project No.60275021).

REFERENCES

Li D.R., Wang S.L., Li D.Y. and Wang X.Z., 2002: Theories and technologies of spatial data knowledge discovery. Geomatics and Information Science of Wuhan University 27(3), 221-233.

Guo Dazhi, Fang Tao, Du Peijun and Yun jiang , 2003: Hierarchical Structure and Scaling for Complex System, Journal of China University of Mining & Technology, Vol.32,No.3, pp:213-217.

Wu, J., 1999: HIERARCHY AND SCALING: EXTRAPOLATING INFORMATION ALONG A SCALING LADDER, Canadian Journal of Remote Sensing Vol.25,No.4,pp.367-380.

R. Agrawal, T. Imielinski, and A. Swami, 1993: Mining Associations Between Sets of Items in Massive Databases, Proceedings of the ACM SIGMOD, Washington, DC. pp. 207-216.

David Wai-Lok Cheung, Vincent T. Ng, Ada Wai-Chee Fu, and Yongjian Fu, 1996: Efficient Mining of Association Rules in Distributed Databases, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 911-922.

Qin Ding, 2002: ASSOCIATION RULE MINING ON REMOTELY SENSED IMAGERY USING P-TREES[D], http: // citeseer.ist.psu.edu/.

Rakesh Agrawal and Ramakrishnan Srikant, 1994: Fast Algorithms for Mining Association Rules in Large Databases, Proceedings of the Twentieth International Conference on Very Large Databases, Santiago, Chile.pp. 487-499.