MINING ASSOCIATION RULES IN GEOGRAPHICAL SPATIO-TEMPORAL DATA

Hong Shu[a*], Xinyan Zhu[b], Shangping Dai[c]

[a, b]National Lab for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Ruoyu Road 129, Wuhan, P. R. China 430079, - shu_hong, zxy@lmars.whu.edu.cn
[c]Department of Computer Science, Central China Normal University, Luoyu Road 152, Wuhan P.R.China 430079 - spdai@mail.ccnu.edu.cn

Commission WG II/2 - Spatial Reasoning, Analysis, and Data Mining

KEY WORDS: Vegetation, Climate, Multi-level, Fuzzy, Association Rules, Spatio-temporal, Data Mining.

ABSTRACT:

For the sake of environmental change monitoring, a huge amount of geospatial and temporal data have been acquired through various networks of monitoring stations. For instance, daily precipitation and air temperature are observed at meteorological stations, and MODIS images are regularly received at satellite ground stations. However, so far these massive raw data from the stations are not fully utilized, or say, geographical spatio-temporal structural information in raw data aren't exposed sufficiently. Upon the requirements of human decision-making, explosive raw data is embarrassed in contrast to starved information or knowledge. This paper makes a short introduction to our research project, i.e., mining association rules in vegetation and climate changing data of northeastern China. We describe a framework of mining association rules in regional vegetation and climate-changing data, in which the methods of Kriging interpolation, wavelets multi-resolution analysis, fuzzy c-means clustering, and Apriori-based logical rules extraction are used respectively. Then, we give the definitions of geographical spatio-temporal transactions and multi-level fuzzy association rules, which play the decisive roles in association rules mining. It is noted that the largest difference between spatial data mining and general data mining is computation of geographical spatio-temporal correlations or variability. The whole procedure of geographical spatio-temporal data mining can be thought of as multi-stage computation of geographical spatio-temporal correlations. At the end of this paper, towards an advanced regional vegetation-climate changing data mining system, several underlying techniques of mining spatio-temporal association rules are initiatively pointed out for next work.

1. INTRODUCTION

As an intensive increase of human activity, nowadays regional and global environments are undergoing dramatic changes. Global climate has obviously shown the trend of getting warmer. In these circumstances, international institutions have jointly made some major plans of earth observation for monitoring environmental changes. Following that, a variety of observation networks have been built worldwide. Also, national governments have constructed numerous geospatial information infrastructures. As a result, these environmental observation networks and geospatial information infrastructures have collected a huge amount of geospatial data. However, these raw data are not utilized to its fullest. Upon the requirements of human decision-making, massive raw data is contradictory to relatively less required information. To overcome this embarrassment, the technology of geospatial data mining is emerging in the geographical information community. Accordingly, academic workers comprehensively make use of applied statistics, machine learning, databases, and information visualization for discovery of geographical knowledge hidden in massive geospatial data (Fayyad U., Piatetsky-Shapiro G., Smyth P., 1996). The concept of geospatial data mining is put forward in the early 90s of last century (Krzysztof Koperski, Jiawei Han, 1995). With reference to general data mining, numerous frameworks for geographical data mining have proposed subsequently over the past decade. However, the technology of geographical data mining, particularly geographical spatio-temporal data mining, is still in its infancy of research. A large number of techniques, such as the formal definition of geographical spatio-temporal transaction, the algorithms of geographical spatio-temporal data conceptualization, and the storage methods of mining data assigned with rich spatio-temporal semantics, remains unsolved so far. In this paper, our study is limited to the technology of mining association rules in regional vegetation and climate changing data. Like classifications and outliers, the association rule is viewed as one typical form of human knowledge.

2. MINING ASSOCIATION RULES IN GEOGRAPHICAL SPATIO-TEMPORAL DATA

Our study area is located in northeastern China, and mainly covers the administrative union of Liaoning province, Jilin province, and Heilongjiang province. Experimental weather observation data, daily precipitation or air temperature, is a time series of 56 years during the period of 1951 to 2006. Daily accumulative precipitation is recorded with a precision of 0.1 mm. Daily average air temperature is recorded with a precision of 0.1 centigrade degree. Experimental vegetation data is a 16-day maximum composite data product of MODIS Normalized Difference Index images. Vegetation data, at the ground resolution of 500m, is a time series of 7 years during the period of 2000 to 2006. Weather observation data is provided by the Meteorological Scientific Data Share and Service Network of China. Vegetation data is freely downloaded at web sites of the MODIS satellite images.

The whole procedure of data mining is completed in three stages, as shown in Figure 1. At the first stage of raw data pre-processing, missing data is filled, and noisy data is filtered in some way. At each certain time, weather observation data, precipitation and air temperature, is interpolated using the method of Kriging. Then, the geographical reference system, e.g., WGS 84, is selected for registration of weather observation data and vegetation data. For the consistency of time periods, weather observation data is temporally truncated into 7 years to match 7 years of vegetation data. Moreover, weather observation data and vegetation data are re-sampled with the same resolution of 16 days in time and 500m in space. At the second stage of data conceptualization, weather observation data and vegetation data are separately clustered by the algorithm of fuzzy c-means clustering (FCM). At the third stage of association rules mining, conceptualized data are flexibly organized into transactions for extracting association rules by the algorithm Apriori (Agrawal R., Imielinski T., Swami A., May 1993). Among Apriori-extracted association rules, many redundant and geographically meaningless rules are eventually filtered with user-specified interest measure thresholds and application-oriented data item constraints. It must be pointed out that, in geographical spatio-temporal data mining, uncertainty is almost involved at each step of data processing, from data pre-processing through data conceptualization until association rules extraction. The final results of data mining are association rules with a comprehensive quality evaluation of interest measures, i.e., support and confidence. Intrinsically, all sources of uncertainty in data mining are reduced to interest measures for final association rules. To enable the usability of mined association rules and meaningful explanations of interest measures for association rules, we make an overall analysis of uncertainty originated from all major steps of data mining.

In nature, discovery of geographical association rules is computation of multivariate spatio-temporal correlations or multivariate spatio-temporal variability through the stages of data mining. Essentially, geographical association rules are the results of multivariate spatio-temporal correlations computation at progressively higher levels. At the first stage of raw data pre-processing, multivariate spatio-temporal correlations are implied by spatio-temporal statistical functions of interpolation and re-sampling. At the second stage of data conceptualization, multivariate spatio-temporal correlations are implied by wavelets multi-resolution decomposition functions and hierarchically fuzzy clustering functions for multivariate spatio-temporal data. At the third stage of association rules mining, multivariate spatio-temporal correlations are implied by Apriori-extracted logical rules. Meanwhile, along with the multi-level representations of knowledge about multivariate spatio-temporal correlations, there exist various sources of knowledge uncertainty induced at each stage of data mining.

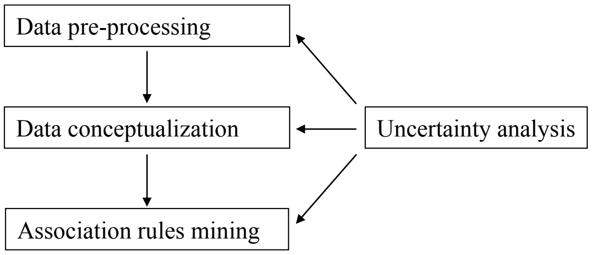

Figure 1. The flow chart of association rules mining

The general algorithm Apriori is applied to geographical spatio-temporal transactions for extracting association rules. Different geographical association rules are mined out of spatio-temporal transactions. Undoubtedly, the organization forms of geographical spatio-temporal transactions are of fundamental significance in data items composition and interest measures for association rules. The next sections are devoted to the explorations of geographical spatio-temporal transactions and the definitions of multi-level fuzzy association rules in the context of regional vegetation and climate changing data.

3. MULTI-LEVEL FUZZY ASSOCIATION RULES

3.1 Hierarchies of Data Conceptualization

The hierarchy is one fundamental property of geographical spatio-temporal structures or multivariate spatio-temporal correlations. Likewise, hierarchically structuralization or conceptualization can be performed for regional vegetation and climate changing data. Moreover, multi-level or generalized association rules can be developed (Strikant R., Agrawal R., 1995).

After spatio-temporal registration of climate and vegetation data with the consistency of spatio-temporal range, resolution, and geographical reference, an integrated geographical spatio-temporal data field is obtained. Considering the strengths of multi-resolution time-frequency analysis of wavelets, wavelet transformations are employed to make a multi-level decomposition of geographical spatio-temporal data.

At each level of wavelets-decomposed data, fuzzy c-means clustering is applied to data respectively at the dimension of space, time, vegetation, air temperature, and precipitation. After that, semantics is assigned to each data cluster with reference to the taxonomy of concepts. In Figure 2, the taxonomy of concepts is generally represented with a directed acycle graph (DAG). For the purpose of conceptual integrity, the root node "All" virtually stands for the top-level concept of the multivariate concepts taxonomy, i.e., All < {Space, Time, Temperature, Precipitation, Vegetation}. At the bottom of the concepts taxonomy, each node stands for numerical or spatio-temporal geometrical data originally exported from wavelet decomposition. The generalization or the hierarchical relation "is a" is algebraically represented with the partial order "<", e.g., Northeastern China < {Heilongjiang, Jilin, Liaoning}. At any level of the concepts taxonomy, concepts can be fuzzy or crisp. For instance, Heat {cold, hot} is a fuzzy concept set, but Province {Heilongjiang, Jilin, Liaoning} is a crisp concept set.

It is seen that, in our project, multi-scale or data scaling is involved twice in data conceptualization. For the first time, multi-scale is interpreted with the multi-resolution of wavelet data decomposition. For the second time, multi-scale is interpreted with the hierarchy of data clustering or conceptualization. Even so, there doesn't exist any redundancy of multi-scale information. In a sense, multi-level data clustering has dialectically illustrated the scale-effects of clustering or conceptualization as shown in the modifiable area units (MAU)

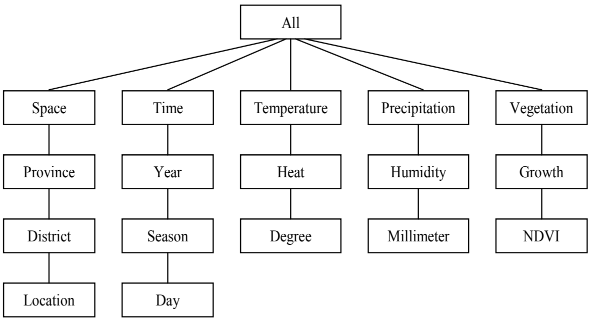

Figure 2. The taxonomy of concepts

3.2 Geographical Transactions and Fuzzy Association Rules

In general, a geographical transaction is a set of geographical data items, each one among which is a data cluster with specific semantics. In particular, a geographical transaction is a geographical object with the characteristics of space, time, air temperature, precipitation, and vegetation. A geographical transaction or a geographical object is assigned with a globally unique identifier, denoted by TID. In this way, a geographical transaction is regarded as a point in the five-dimensional feature space. Space and time are two independent data items as other three independent data items of air temperature, precipitation, and vegetation. In geographical transactions, no spatio-temporal relationships are explicitly represented, and all spatio-temporal relationships are implicitly represented with discretely distributed values of space and time features.

In cognitive sciences, it is proved that the hierarchy is one property of human mental models. Also, fuzziness is a basic property of human language. Furthermore, fuzzy logic is one law of human thoughts. Hence, it is nature that knowledge, discovered from databases or transaction sets, is represented with logical rules of fuzzy predicates or fuzzy data items. A general association rule, implication of the form X=>Y, is constituted of crisp data items. Accordingly, a fuzzy association rule is constituted of fuzzy data items. It is apparent that the general association rule is a special case of the fuzzy association rule. Certainly, interest measures for the fuzzy association rule are evaluated more complicatedly than those of the general association rule.

TID	Space			Time				Temperature		Precipitation		Vegetation	
	Heilongjiang	Jilin	Liaoning	Spring	Summer	Autumn	Winter	Cold	Hot	Dry	Wet	Poor	Rich
t1	$\mu_{Heilongjiang}$(t1(Space))	μ_{Jilin}(t1(Space))	$\mu_{Liaoning}$(t1(Space))	μ_{Spring}(t1(Time))	μ_{Summer}(t1(Time))	μ_{Autumn}(t1(Time))	μ_{Winter}(t1(Time))	μ_{Cold}(t1(Temperature))	μ_{Hot}(t1(Temperature))	μ_{Dry}(t1(Precipitation))	μ_{Wet}(t1(Precipitation))	μ_{Poor}(t1(Vegetation))	μ_{Rich}(t1(Vegetation))
…	…	…	…	…	…	…	…	…	…	…	…	…	…
tn	$\mu_{Heilongjiang}$(tn(Space))	μ_{Jilin}(tn(Space))	$\mu_{Liaoning}$(tn(Space))	μ_{Spring}(tn(Time))	μ_{Summer}(tn(Time))	μ_{Autumn}(tn(Time))	μ_{Winter}(tn(Time))	μ_{Cold}(tn(Temperature))	μ_{Hot}(tn(Temperature))	μ_{Dry}(tn(Precipitation))	μ_{Wet}(tn(Precipitation))	μ_{Poor}(tn(Vegetation))	μ_{Rich}(tn(Vegetation))

Table 1. Geographical transactions constituted of fuzzy data items

Formally, geographical transactions and fuzzy association rules are defined as follows.

Definition 1 (Geographical transaction). A geographical transaction set is a set of geographical transactions, denoted by D={t}.

Definition 2 (Fuzzy geographical transaction). Given a set of attributes, denoted by {A}. For each attribute A, there exists a set of fuzzy concepts, denoted by {a}. The whole set of linguistic-valued data items is denoted by I={<A, a>}. A geographical transaction, denoted by t, is composed of geographical data items, which are taken from the whole set of linguistic-valued data items. $\mu_a(t(A))$ stands for the membership grade of fuzzy concept a of attribute A for tuple t, or the membership grade of data item a for transaction t. A geographical transaction constituted of fuzzy data items is named fuzzy geographical transaction.

In our study, the whole set of linguistic-valued data items, I, is temporally realized with {<Space, Heilongjiang>, <Space, Jilin>, <Space, Liaoning>, <Time, Spring>, <Time, Summer>, <Time, Autumn>, <Time, Winter>, <Temperature, Cold>, <Temperature, Hot>, <Precipitation, Dry>, <Precipitation, Wet>, <Vegetation, Poor>, <Precipitation, Rich>}. In example of fuzzy geographical transaction set for Apriori-based fuzzy association rules extraction is presented in Table 1.

Definition 3 (Fuzzy association rule). A fuzzy association rule is an implication of the form, X=>Y, with interest measures of support and confidence. In the rule, antecedent X and consequent Y are two nonempty subsets of the whole set of linguistic-valued data items, denoted by $X, Y \subseteq I$. Meanwhile, no data items of an association rule exist simultaneously in antecedent X and consequent Y, denoted by $X \cap Y = \varnothing$.

Definition 4 (Interest measures for the fuzzy association rule). Interest measures for an association rule are referred to as support and confidence, denoted by support(X=>Y) and confidence(X=>Y). The formulas of interest measures are

$$\text{support}(X \Rightarrow Y) = \frac{\sum\limits_{t \in D < A,a > \in X \cup Y} \min \{\mu_a(t(A))\}}{|D|} \quad (1)$$

$$\text{confidence}(X \Rightarrow Y) = \frac{\sum\limits_{t \in D < A,a > \in X \cup Y} \min \{\mu_a(t(A))\}}{\sum\limits_{t \in D < A,a > \in X} \min \{\mu_a(t(A))\}} \quad (2)$$

Note that the most common choice for the t-norm is minimum as given above, yet the product has also been applied as follows (Dubois Didier, Hüllermeier Eyke, Prade Henr, 2006).

$$\text{support}(X \Rightarrow Y) = \frac{\sum\limits_{t \in D < A,a > \in X \cup Y} \Pi \ \mu_a(t(A))}{|D|} \quad (3)$$

$$\text{confidence}(X \Rightarrow Y) = \frac{\sum\limits_{t \in D < A,a > \in X \cup Y} \Pi \ \mu_a(t(A))}{\sum\limits_{t \in D < A,a > \in X} \Pi \ \mu_a(t(A))} \quad (4)$$

As for semantics, the support quantitatively implies the extrinsic importance or occurrence frequency of an association rule, and the confidence quantitatively implies the intrinsic correctness or rationality of an association rule. Larger the support, more important the association rule is. Larger the confidence, more correct the association rule is. Given definitions of multi-level fuzzy association rules and interest measures, the algorithm Apriori is easily applied to fuzzy geographical transactions for extracting fuzzy association rules, which have interest measures more than user-specified minimum support and confidence.

4. CONCLUSION

In the context of our project, the framework of mining multi-dimensional association rules in regional vegetation and climate-changing data is specially designed. Basically, the framework consists of data pre-processing, data conceptualization, association rules mining, and overall uncertainty analysis. It is recognized that the difference between general data mining and geographical data mining is the representation and computation of geographical spatio-temporal correlations. In our study, geographical spatio-temporal correlations are evaluated respectively with the methods of geostatistics interpolation, wavelet data decomposition, fuzzy c-means clustering, and Apriori-based logical rules extraction.

To some extent, data mining is often known as knowledge discovery. The results of data mining must be easily understandable for humans. Undoubtedly, the association rule, a specific case of the logical rule, is one typical form of human knowledge. Commonly, there exist the properties of hierarchy and fuzziness in geographical spatio-temporal structures. In our study, the hierarchy is realized through multi-scale data decomposition based on wavelets transformation, and multi-level data clustering with reference to a taxonomy of concepts. The fuzziness is realized through fuzzy c-means clustering and interest measures for fuzzy association rules.

Since geographical association rules are mined out of geographical transactions, construction of geographical transactions is of large significance. At present, we only consider a geographical transaction as a composition of features of space, time, air temperature, precipitation, and vegetation, or as a five-dimensional geographical object. This organization of transactions is clearly weak in spatio-temporal semantics modeling. Besides, here association rules mining is only performed within the transaction. This has put strict constraints on the variety of association rules.

In the future, while enhancing data processing functions in the framework, several underlying technical issues are put on the prior schedule. These critical technical issues are: 1) how to flexibly organize geographical transactions for enriching geographical spatio-temporal correlations or semantics, 2) how to extend intra-transaction data mining to inter-transaction data mining, and 3) how to mine association rules constituted of across-level concepts in the concept hierarchy, particularly constituted of concepts in the fuzzy hierarchy of concepts.

REFERENCES

[1] Agrawal R., Imielinski T., Swami A., May 1993. Mining association rules between sets of items in large databases. In *Proceedings of ACM SIGMOD*, pp. 207-216.

[2] Dubois Didier, Hüllermeier Eyke, Prade Henr, 2006. A systematic approach to the assessment of fuzzy association rules. *Data Mining and Knowledge Discovery*, 13(2), pp.167-192.

[3] Fayyad U., Piatetsky-Shapiro G., Smyth P., 1996. From data mining to knowledge discovery: an overview. Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., Editors, *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, pp.83-115.

[4] Krzysztof Koperski, Jiawei Han, 1995. Discovery of spatial association rules in geographic information databases. In *Proceedings of the SSD*, pp.47-66.

[5] Strikant R., Agrawal R., 1995. Mining generalized association rules. In *Proceedings of the 21st International Conference on Very Large Databases VLDB Conference*, Zurich, Switzerland, pp.402～419.

ACKNOWLEDGEMENTS

This research is funded jointly by National Grand Fundamental Research Program of China (No. 2006CB701305) and National High Technology Research and Development Program of China (863 Program). The preliminary experimental tests made by the student members of our research group in geographical spatio-temporal data mining are gratefully acknowledged.