Mining Association Rules In Spatio-Temporal Data

Jeremy Mennis and Junwei Liu

Department of Geography, University of Colorado,
Boulder, CO, 80309, U.S.A.,
Tel (303) 492-4794;
Fax (303) 492-7501;
Email jeremy@colorado.edu

Department of Geography, University of Colorado,
Boulder, CO, 80309, U.S.A.,
Fax (303) 492-7501;
Email Junwei.Liu @colorado.edu

Biography

Jeremy Mennis (Ph.D. 2001, The Pennsylvania State University) is an Assistant Professor of Geography at the University of Colorado. His research interests are in spatio-temporal databases and geographic data mining and knowledge discovery. He has published in journals such as International Journal of Geographical Information Science, International Journal of Remote Sensing, and The Professional Geographer.

Introduction

Spatial data mining is an emerging research area dedicated to the development and application of novel, typically inductive, computational techniques for the analysis of very large, heterogeneous spatial databases (Miller and Han 2001). There has been relatively little research on adapting classical data mining techniques to data with both a spatial and temporal component. The purpose of this research is to demonstrate the application of a certain type of data mining technique, association rule mining, to spatio-temporal data. As a case study, we use association rule mining to explore the spatial and temporal relationships among geographic data that characterize socioeconomic and urban land cover change in the Denver metropolitan area, Colorado, U.S.A. Strategies for data preprocessing to support spatio-temporal association rule mining are discussed.

Assocation Rule Mining

Association rule mining seeks to discover associations among transactions within relational databases (Agrawal et al. 1993). An association rule takes the form *A* ? *B* where *A* (the antecedent) and *B* (consequent) are sets of predicates. Association rule mining uses the concepts of support and confidence. The support is the probability of a record in the database satisfying the set of predicates contained in the antecedent. The confidence is the ratio of the probability that a record in the database satisfies both antecedent and consequents sets of predicates to the support. The support and confidence of a rule are reported in parentheses following the rule, i.e. '(support%, confidence%)'.

By setting minimum support and confidence thresholds, one may use association rule mining to identify only *strong* rules. A spatial (spatio-temporal) association rule contains a spatial (spatio-temporal) relationship predicate in the antecedent or consequent of the rule (Koperski and Han 1995).

Data And Methods

As a case study of spatio-temporal association rule mining, we focused on the analysis of socioeconomic and urban land cover change data covering the Denver, Colorado, U.S.A. region from 1980 to 2000. Data indicating race, educational attainment, employment, and income for 1980 and 2000 were acquired from the U.S. Bureau of the Census and aggregated to the Census 2000 tract level. Vector land cover data for 1980 and 2000, generated from historic aerial photography, were acquired from the U.S. Geological Survey (USGS). Data on limited and unlimited access highways were acquired from the Environmental Systems Research Institute, Inc. (ESRI) Streets Database.

One challenge in spatial data mining is the performance/data storage tradeoff between preprocessing spatial relationships among geographic objects and computing those relationships on-the-fly (Klosgen and May 2002). A number of approaches have been developed to address this issue, including the use of R* trees and minimum bounding rectangles for fast computation of spatial relationships (Kopersky and Han 1995), the use of spatial relationship indices (Ester et al. 2000), and the encoding of spatial relationships among certain target sets of geographic objects prior to data mining (Malerba et al. 2002).

We address this challenge in our case study by preprocessing spatial relationships through data integration so that we produce a single relation (table in the relational database) which can be mined using shareware non-spatial association rule mining software. In this relation, each record represents a Census 2000 tract and all variables (e.g. land cover, distance to highway) are mapped to the tract level. In order to capture the spatial coincidence between land cover and socioeconomic status at a particular year, the percent of each land cover type was calculated for each tract for 1980 and 2000. Each tract was also attributed with fields describing the 1980minimum distances to limited and unlimited highways, and to the nearest areas of residential, commercial, and industrial land covers. To support the mining of change through time, changes from 1980 to 2000 were calculated for each tract, e.g. the change in percent developed land and the change in educational attainment. Because association rule mining works with nominal and ordinal data, all interval/ratio variables (e.g. percentages, distances) were transformed into ordinal data through a ten class quantile classification. We mined this relation using the association rule mining package CBA (Classification Based on Associations) (Liu et al. 1998).

Results

Preliminary results indicate that the majority of the strongest rules (those with the highest confidence values) are fairly intuitive. For example, one rule states that tracts that are between 97% and 100% developed in 1980 decrease in percent developed area (from 1980 to 2000) (19%, 79%). Other rules are more interesting and reveal patterns among

development and socioeconomic status. An example of such a rule is: tracts that are between 85% and 92% developed in 1980 and in which the change in percent minority is greater than 31% have an increase in percent living below the poverty line greater than 7% (2%, 71%). In other words, highly developed tracts in 1980 that undergo a large increase in minority status also tend to become increasingly poor. On the other hand, another rule states that tracts that are between 10% and 27% developed in 1980 and in which the change in percent minority is between 11% and 15% have a decrease in percent living below the poverty line (1%, 100%).

It is important to note that the support for the previous two rules (2% and 1%, respectively) represents a small number of tracts, seven and five tracts out of 454, respectively. However, the ten class quantile classification implies that each class will contain only approximately 10% of all tracts. Thus any antecedent with multiple predicates is ensured of having a support of less than 10%, and likely much less since the support captures only those records that contain all the predicates stated in the antecedent.

Conclusion

This study demonstrates how a particular data mining technique, association rule mining, can be applied to spatio-temporal data. We have also demonstrated how preprocessing spatial and temporal relationships aids in association rule mining and allows for the spatio-temporal application of non-spatio-temporal association rule mining software. For the sake of efficiency, we have chosen certain options in terms of data classification and the derivation of certain spatial and temporal relationships. Future research should focus on the impacts of using different data classification schemes (e.g. quantile versus equal interval) and strategies for capturing spatial relationships (e.g. metric versus topologic).

References

Agrawal, R., Imielinski, T. and Swami, A. (1993) Mining association rules between sets of items in large databases, in P. Buneman and S. Jojodia (Eds), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207-216, Washington D.C.: ACM Press.

Ester, M., Frommelt, A., Kriegel, H.-P. and Sander, J. (2000) Spatial data mining: database primitives, algorithms and efficient DBMS support, *Data Mining and Knowledge Discovery* 4, 193-216.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) The KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM* 39(11), 27-34.

Klosgen, W. and May, M. (2002) Spatio-temporal subgroup discovery, in R. Ladner, K. Shaw and M. Abdelguerfi (Eds), *Mining Spatio-Temporal Information Systems*, 149-168. Boston: Kluwer Academic Publishers.

Koperski, K. and Han, J. (1995) Discovery of spatial association rules in geographic information databases, in M.J. Egenhofer and J.R. Herring (Eds), *Advances in Spatial Databases, 4th International Symposium, SSD'95*, 47-66, Berlin: Springer.

Liu, B, Hsu, W. and Ma, W. (1998) Integrating classification and association rule mining, in R. Agrawal and P. Stolorz (Eds), *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, 80-86, Menlo Park, California: AAAI.

Malerba, D., Lisi, F.A., Appice, A. and Sblendorio, F. (2002) Mining spatial association rules in census data: a relational approach, in P. Brito and D. Malerba (Eds), *Notes of the ECML/PKDD 2002 Workshop on Mining Official Data*, 80-93, Helsinki: Helsinki University Printing House.

Miller, H.J. and Han, J. (2001) Geographic data mining and knowledge discovery: an overview, in H.J. Miller and J. Han (Eds), *Geographic Data Mining and Knowledge Discover*, 3-32, London: Taylor and Francis.