

Visualization and Database Support for Geographic Meta-Mining

Jeremy Mennis

Department of Geography, UCB 260, University of Colorado, Boulder, CO 80309

Phone: (303) 492-4794, Fax: (303) 492-7501, Email: jeremy@colorado.edu

White Paper:

University Consortium for Geographic Information Science (UCGIS) Visualization and Geographic Knowledge Discovery Workshop, Lansdowne, Virginia, November 18-20, 2003

Introduction

Geographic data mining can be defined as a set of exploratory computational and statistical approaches for analyzing very large spatial and spatiotemporal data sets. Data mining techniques are often grouped into categories that include clustering, categorization, summarization, rule-mining, and feature extraction. All of these types of techniques are generally oriented towards identifying spatial or spatio-temporal patterns in geographic observations or measurements. The identification of these patterns is intended to spur hypothesis generation as to the geographic processes from which the patterns are generated.

Data mining can be considered one step in the larger process of geographic knowledge discovery (Fayyad et al. 1996). This process is both interactive and iterative and includes steps such as data selection, data cleaning, and the interpretation of data mining results. While there are a variety of academic and commercial data mining software available, there are few comprehensive knowledge discovery software environments. The process of knowledge discovery is supported wholly by the analyst, who is responsible for keeping track of the results of multiple data mining ‘runs,’ for instance descriptions of rules or extracted features. These results are ideally elements of knowledge – summaries of patterns embedded within the observational data. However, these results can also be considered a form of data themselves that are often in need of further analysis to yield useful interpretation. In many cases, the data sets resulting from data mining are large and complex, yet there are few computational techniques for managing these data mining results to fully support the knowledge discovery process.

The analysis of the results of data mining is called *meta-mining* (Abraham and Roddick 1999). I argue here that geographic knowledge discovery software demands support not only for data mining but also for the ability to visually and algorithmically meta-mine the results of data mining in an interactive and iterative manner. Computational support for meta-mining is dependent on the visualization and database representation of the rules, clusters, and features that are the results of data mining. Thus, geographic knowledge discovery software environments must incorporate visualization and semantic database modeling strategies for the representation of, and interaction with, these knowledge elements.

Database Support for Geographic Meta-Mining

There has been substantial research in computer and information science, artificial intelligence, and related fields in computational knowledge representation and semantic data models. This research has been extended by geographers and others for geographic databases. Recently, much of this research has come under the heading of *ontology* – the development and formal encoding of the conceptual elements and relationships composing particular application domains. Only a handful of these research efforts have been directed towards geographic meta-mining, however.

There has been some research in meta-mining association rules. Association rule mining identifies rules in transactional databases based on the rate of co-occurrence of certain attributes within a set of transactions. Spatial association rule mining is defined as when a rule contains a spatial relationship. A data set with just a handful of attributes, each with a just a few potential values, can generate tens of thousands of rules. Unlike in a formal statistical analysis, there is no measure of significance in association rule mining that one may use to cull these results to a manageable (and interpretable) size. Rather, results may be evaluated by assigning a measure of ‘interestingness’ to each rule, such as the confidence and support metrics. Interestingness measures are fairly subjective devices, however, and in my own experience I have found that rules that are of actual interest may score as relatively ‘uninteresting,’ putting in doubt the utility of such metrics.

Researchers have suggested ways of facilitating the search for interesting association rules (i.e. meta-mining rule sets) through database support for encoding information *about* the mined rules. One approach is to design a rule ‘template,’ a general type of rule that is of particular interest and which may be extracted from the complete set of mined rules for closer examination (Fu and Han 1995). Another approach for spatio-temporal association rule mining is to look at the changes in the interestingness measures of a particular rule over time (Spiliopoulou and Roddick 2000). These approaches demand that something beyond the basic transactional data be encoded within a database. In the rule template approach, knowledge in the form of a pattern (i.e. a general type of rule) must be encoded so that rules that fit this pattern may be identified. Analyzing changes in rules over time implies that the rules and their attributes (e.g. interestingness measures) derived for each time period (or moment) be encoded to facilitate temporal comparison.

These principles may be extended to other types of geographic data mining. For example, in feature extraction the features that are identified may be encoded as individual database objects with their own set of attributes. In addition, the ‘rules’ (in whatever form) used to extract the features from the observational data may also be encoded in the database. The feature attributes may then be meta-mined using statistics, data mining, or visualization (Mennis and Peuquet 2003). The analyst then has the option of recalling the feature extraction rules from the database, refining those rules, and rerunning the initial feature extraction to derive a new set of features and associated attributes, which may then be meta-mined, and so on in an iterative process.

Geographic meta-mining demands geographic data structures that can encode not only data but also knowledge. This knowledge may take the form of the patterns or rules derived from data mining, the choice and parameterization of the data mining algorithm used to extract a particular data mining result, or contextual knowledge of the application

domain under investigation that the analyst brings to the analysis (e.g. a rule template for use in association rule mining). In order to be useful, the analyst must be able to interact with these knowledge elements. In terms of the results of data mining, the analyst should be able to review, summarize, or apply further analysis to these results. The analyst should also be able to easily retrieve and view the choice and parameterization of a data mining algorithm associated with any result set. Finally, the analyst should be able to easily declare domain knowledge in a variety of forms, for example as a rule template in association rule mining or as an exemplar feature with which a feature extraction algorithm can be used to find similar features.

Visualization Support for Geographic Meta-Mining

There are many established approaches for visualizing geographic data: maps, scatterplot matrices, and parallel coordinate plots, just to name a few. Techniques have also been developed for visualizing spatio-temporal data, such as animation and small multiples. Other visual data mining techniques have been developed for large, multidimensional data sets, potentially spatial, and are often categorized as geometric, icon-based, hierarchical, and pixel-based approaches. These techniques typically are used for spatial data in which multiple attributes are recorded for each location, or each location at multiple times. A hallmark of exploratory visualization is the ability to interact with multiple visual representations through brushing and linked displays.

Visualization techniques to support geographic meta-mining may in some cases be easily transferred from those used for visualizing geographic data. If the results of geographic data mining take the form of a set of individual georeferenced observations with multiple attributes attached, then the ‘conventional’ visualization strategies described above can be easily applied. For example, if feature extraction is applied to a spatial data set and yields a set of geographic features, each with its own set of attributes, these features can be mapped or explored using a scatterplot matrix.

If, however, the results of data mining take a form that is quite different from the standard location/attribute set format of most geographic data, such as a set of association rules, new kinds of specialized visualization techniques must be developed. A simple approach for visualizing the results of association rule mining would be to simply map where certain rules occur. A more novel approach might use spatialization (the use of a spatial metaphor to ‘map’ non-spatial data values) to display the different rules in an attribute space where the dimensions consist of different attributes and interestingness metrics.

A key component of visual meta-mining is interactivity and the connection with database knowledge representation. For example, consider a hypothetical situation in which a feature extraction algorithm has been used to identify a set of individual features embedded in the data. These features may be mapped and plotted on a scatterplot matrix. The analyst may identify one particular feature that is of interest and want to find other features like that one in the data. A meta-mining system should support the ability of the analyst to select a particular feature from a visualization, encode that feature in the database as a knowledge element, and use that knowledge to find similar features in the database.

Conclusions and Challenges

A brief search on the web yields many individual data mining software tools. Some of these tools support, or may be adapted for, spatial and spatiotemporal data. However, the computational support for the *process* of geographic knowledge discovery is generally lacking. This is due to the fact that the management of the knowledge discovery process (i.e. keeping track of, data mining choices and parameterizations, results, and interpretations) must be maintained ad hoc by the analyst. Advances in developing knowledge discovery software environments should seek not only to incorporate a variety of data mining tools, but also the means to support geographic meta-mining and, consequently, the interactive and iterative nature of the knowledge discovery process.

Below are five research challenges for developing such knowledge discovery environments:

1. Development of data structures to support the storage of both data and knowledge (i.e. data mining results as well as analyst-defined domain knowledge), and the mapping from one to the other
2. Development of modes of interaction for database representations of knowledge
3. Development of visualization techniques for the results of data mining
4. Development of modes of interaction for the visualization of data mining results
5. Development of the linkage among visual meta-mining techniques and data structures for knowledge representation

References

- Abraham T and Roddick J F 1999 Incremental meta-mining from large temporal data sets. In Kambayashi Y, Lee D K, Lim E-P, Mohania M, and Masunaga Y (eds) *Advances in Database Technologies: Proceedings of the First International Workshop on Data Warehousing and Data Mining*. Berlin, Springer Verlag: 41-54
- Fayyad U, Piatetsky-Shapiro G, and Smyth P 1996 From data mining to knowledge discovery: an overview. In Fayyad U, Piatetsky-Shapiro G, Smyth P, and Uthurusamy R (eds) *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA, AAAI/MIT Press: 1-34
- Fu Y and Han J 1995 Meta-rule-guided mining of association rules in relational databases. In *Proceedings of the International Workshop on the Integration of Knowledge Discovery with Deductive and Object-Oriented Databases*: 39-46
- Mennis J and Peuquet D J 2003 The role of knowledge representation in geographic knowledge discovery. *Transactions in GIS* 7: 371-391
- Spiliopoulou M and Roddick J F 2000 Higher order mining: modeling and mining the results of knowledge discovery. In Ebecken N and Brebbia C A (eds) *Data Mining 2000 – Proceedings of the Second International Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields*. Southampton, UK, WIT Press: 309-20