Mining Spatial Data Using An Interactive Rule-Based Approach

Donna J. Peuquet and Diansheng Guo
Department of Geography, Pennsylvania State University

With the advent of very large spatial databases, it is beyond human capacity to examine and understand the information contained within such volumes of data directly. Although data mining has been recognized as a key means of finding patterns in large databases, general data mining methods alone are not sufficient for *spatial* data mining. The strengths of the computer’s ability to perform numerical calculations on large volumes of data with mechanical speed and accuracy needs to be linked with the human’s ability to visually recognize complex patterns very quickly using expertise and intuition. The research reported upon here investigates how interactive visualization techniques can be integrated with rule-based methodologies in order to provide a cooperative, and thereby powerful, data mining and knowledge discovery environment for very large geographic databases.

Besides simply visualizing the results of data mining procedures, geographic visualization (GVis) capabilities need to become an integral component of the data mining process. However, visualization software currently does not allow interactive access to database and data mining packages. In the current paper, we present the architecture for a prototype example of a combined, interactive data mining and visualization environment, including a conceptual data model for representing and efficiently manipulating space-time data. We focus our discussion on how to integrate the C4.5 algorithm (Quinlan, 1993), as the first of a suite of data mining algorithms, with geographic visualization capabilities.

The C4.5 decision tree method, one of the most widely used methods for inductive learning, can discover relationships between the classification of objects and their attributes. As an attribute-oriented machine learning method, C4.5 is targeted to non-spatial business data in tabular format. To fully exploit the existing capability of the
algorithm for spatial data mining, it is essential to transform necessary spatial properties into attributes, including those are implicit in spatial database such as distance and neighbors. Our implemented data model, utilizing the POET object-oriented DBMS, supports the creation, storage and retrieval of these attributes as well as other non-spatial attributes.

Relationships lurking in spatial data are often localized, both in geographical space (Openshaw, 1999) and attribute space. It is more effective to focus on a subset of the data to find interesting local patterns. Combining visualization with C4.5 facilitates this process. If the whole dataset is input into C4.5 data algorithm, some local patterns tend to be “drowned out” (Simoudis, 1996). We use the visualization module as the control interface for C4.5 module and guide it to discover localized patterns. Moreover, by visualization, the user can recognize more complex local patterns than the C4.5 can identify, such as linear or non-linear relationships among several attributes. Then, some attributes may be removed and new attributes may be added to achieve a better result in C4.5 module. In addition, analyzing the decision tree result, or, letting the algorithm process different sub-datasets rather than processing the data in monolithic fashion, the user can obtain information on where the possible bottleneck is and what to focus on in the visualization module.

