Discriminating Patterns for Surveying and Mapping Geospatial Relationships

Wei Ding
wei.ding@umb.edu
http://www.cs.umb.edu/~ding/
Computer Science Department
University of Massachusetts Boston

Many geospatial systems comprises complex, nonlinear, and spatially non-uniform relationships making it difficult to formulate a model, either deterministic or statistical, linking a response variable to its predictors. We propose a methodology for surveying and mapping all spatially distributed predictors-response relationships present in a given data set. The method aims at the discovery of geographical regionalization of such relationships, and, in particular, at revealing diversity of complexes of predictors associated with a particular level of response. Thus, whereas the bulk of existing work on intelligent data analysis of geospatial data concentrates on regression and classification, the focus of this work is on automatic inventorization of relationships. An input is a geospatial data set where objects are the pixels carrying local values of predictors and response variables. The method, based on association analysis technique of identifying emerging patterns, finds all the patterns of predictors that are frequent in a region characterized by a given level of response but absent or infrequent outside that region. An output is a list of spatial footprints, each characterized by a unique "controlling patterns" —a list of predictors' complexes that locally correlate with a given level of response variable. Mapping the spatial extents of footprints corresponding to various controlling patterns reveals geographic regionalization of predictors-response relationships.

The method consists of a number of stages: (1) integration of geospatial data into the format suitable to association analysis, (2) determining a desire level of response, (3) calculating controlling patterns and their footprints, and (4) summarizing controlling patterns into the products which are of immediate use to domain experts.

Stage (1) utilizes a data categorization procedure that accommodates non-Gaussian distributions and assures that all variables are categorized in a correspondent manner. Stage (2) uses a value-iteration method introduced to optimally split the spatial domain of response variable into the level of interest and the rest. Stage (3) relies on mining for frequent discriminating patterns. Stage (4) uses a new similarity measure between patterns of ordinal features to summarize controlling patterns calculated in Stage (3).

The method is applied to a real-world geospatial data set pertaining to the high density of vegetation within the continental United States. The response is the NDVI measurement (proxy for density of vegetation) and the predictors are eight variables related to various aspects of climate, soil, and topography. Resultant controlling patterns of high vegetation are clustered into six super-patterns using a pattern similarity measure. These six patterns and their associated footprints represent six different regimes of high vegetation across the United States.

Speaker Bio: Wei Ding is an Assistant Professor of Computer Science at the University of Massachusetts Boston. She received her M.Sc. degree in Software Engineering from the George Mason University in 2000 and her Ph.D. degree in Computer Science from the University of Houston in 2008. She has 7-year teaching experience in Computer Science and 8-year industrial working experience in banking, software development, and web technology. Her main research interests include Data Mining, Machine Learning, Artificial Intelligence, Computational Semantics, and with applications to astronomy, geosciences, and environmental sciences. Her research projects are currently sponsored by NASA.