Exploring Labeled Spatial Datasets Using Association Analysis (Demo Paper)

Tomasz F. Stepinski Josue Salazar
Lunar and Planetary Institute
{tom, salazar}@lpi.usra.edu

Wei Ding
University of Massachusetts Boston
ding@cs.umb.edu

ABSTRACT

We use an association analysis-based strategy for exploration of multi-attribute spatial datasets possessing naturally arising classification. In this demonstration, we present a prototype system, ESTATE (Exploring Spatial daTa Association patTErns), inverting such classification by interpreting different classes found in the dataset in terms of sets of discriminative patterns of its attributes. The system consists of several core components including discriminative data mining, similarity between transactional patterns, and visualization. An algorithm for calculating similarity measure between patterns is the major original contribution that facilitates summarization of discovered information and makes the entire framework practical for real life applications. We demonstrate two applications of ESTATE in the domains of ecology and sociology. The ecology application is to discover the associations of between environmental factors and the spatial distribution of biodiversity across the contiguous United States, and the sociology application aims to discover different spatio-social motifs of support for Barack Obama in the 2008 presidential election.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: [Design Methodology][Pattern analysis]; H.2.8 **[Database Management**]: Database Application—spatial databases and GIS

Keywords

association patterns, similarity measure, summarization, visualization, biodiversity, political analysis

1. INTRODUCTION

Advances in gathering spatial data and progress in Geographical Information Science (GIS) allow domain experts to monitor complex spatial systems in a quantitative fashion leading to collections of large, multi-attribute datasets. The complexity of such datasets hides domain knowledge that may be revealed through systematic exploration of the overall structure of the dataset. Often, datasets of interest either possess naturally present classification, or the classification is apparent from the character of the dataset and can be performed without resorting to machine learning. The purpose of this paper is to introduce a strategy for thorough exploration of such datasets. The goal is to discover all combinations of attributes that distinguish between the class of interest and the other classes in the dataset. The proposed strategy (ESTATE) is a tool for finding explanation and/or interpretations behind divisions that are observed in the dataset. Note that the aim of ESTATE is the reverse to the aim of classification/prediction tools; whereas a classifier starts from attributes of individual objects and outputs classes and their spatial extents, the ESTATE starts from the classes and their spatial extents and outputs the concise description of attribute patterns that best define the individuality of each class.

ACM SIGSPATIAL GIS'10, Nov 2–5, 2010, San Jose, CA, USA.

2. ESTATE FRAMEWORK

The ESTATE prototype system is built upon our research work in [1-5]. The strategy is underpinned by the framework of association analysis that assures that complex interactions between all attributes are accounted for in a model-free fashion. Specifically, we rely on the contrast data mining, a technique for identification of discriminative patterns─ associative itemsets of attributes that are found frequently in the part of the dataset affiliated with the focus class but not in the remainder of the dataset. A collection of all discriminative patterns provides an exhaustive set of attribute dependencies found only in the focus class. These dependencies are interpreted as knowledge revealing what sets the focus class apart from the other classes. The set of dependencies for all classes is used to explain the divisions observed in the dataset.

3. DEMONSTRATION

We demonstrate two applications of ESTATE in the domains of ecology and sociology. Fig. 1 depicts the associations between 32 environmental factors and the spatial distribution of biodiversity [2]. The data set has 21,039 data carrying pixels of bird diversity data. ESTATE reports 5 pattern clusters of 1,503 identified patterns that discriminate high-biodiversity from low-biodiversity. Fig. 1A illustrates the footprints of the 5 pattern clusters, and Fig. 1B shows the bar-coded description for the 5 clusters corresponding to different biodiversity regimes. If a given category is absent within a cluster the bar is gray; black bars with increasing thickness denote categories with increasingly large presence in a cluster. 5 clusters indicate 5 distinct motifs of environmental attributes associated with high levels of biodiversity. The results can help develop optimal strategy for protecting bird species given limited resources.

Fig. 2 shows the result of 2008 presidential election data for 3,108 counties located within the contiguous United States using 13 socio-economic indicators. We identify 4 super patterns (clusters of patterns) out of 3,097 patterns that discriminate counties won by Obama from those won by McCain. Fig. 2A is the Sammon's map that visualizes in 2-D the "distance" between the patterns—similar patterns are close to each other on the map. Fig. 2B shows 4 different colors of points corresponding to patterns on the map

represents 4 clusters found using the agglomerative clustering. Fig. 2C shows geographical distribution of footprints corresponding to the 4 super patterns. Fig. 2D indicates the spatio-social motifs of electoral support for Obama in terms of socio-economic indicators.

4. CONCLUSIONS

We have developed the ESTATE framework in order to understand naturally occurring divisions in terms of dataset attributes. In a broader sense, the purpose of ESTATE is reverse to the purpose of a classification. A crucial component of the ESTATE is the pattern similarity measure that enables clustering of similar patterns into agglomerates. ESTATE offers a model-free alternative to approaches based on regression. Successfully application to two real world case studies in different domains indicate broad application appeal of the proposed framework.

5. ACKNOWLEDGMENTS

This work was partially supported by NSF Grant IIS-0812271.

6. REFERENCES

[1] T. Stepinski, J. Salazar, W. Ding. "Discovering Spatio-Social Motifs of Electoral Support Using Discriminative Pattern Mining." COM.geo. 2010.

[2] T. Stepinski, J. Salazar, W. Ding, D. White. "ESTATE: Strategy for Exploring Labeled Spatial Datasets Using Association Analysis." DS10. 2010.

[3] T. Stepinski, W. Ding, C. Eick. "Controlling Patterns of Geospatial Phenomena." GeoInformatica (2010).

[4] T. Stepinski, W. Ding, C. Eick. "Discovering Controlling Factors of Geospatial Variables." ACM SIGSPATIAL GIS. 2008.

[5] W. Ding, T. F. Stepinski, J. Salazar. "Discovery of geospatial discriminating patterns from remote sensing datasets." SDM. 2009.

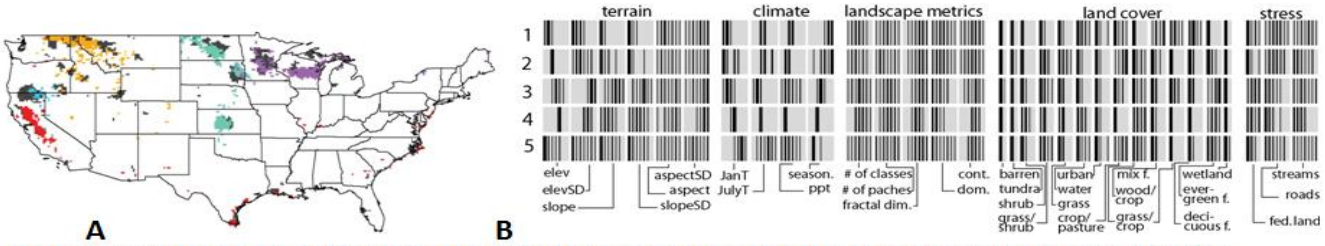

Fig. 1. (A) Spatial footprints of five pattern clusters of bird biodiversity. White: not high biodiversity region; gray: high biodiversity region; purple (cluster #1), light green (cluster #2), yellow (cluster #3), blue (cluster #4), and red (cluster #5). (B) Bar-code representation of the five regimes (clusters) of high biodiversity.

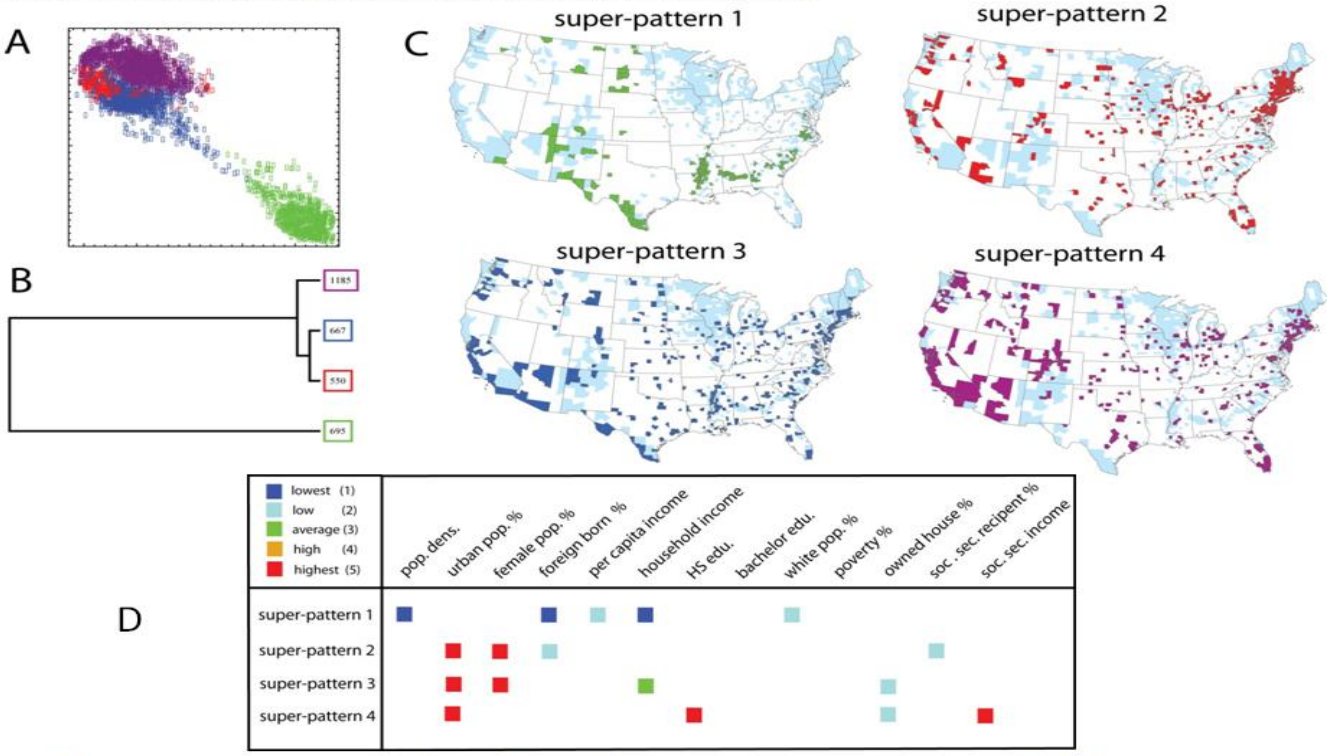

Fig. 2. (A) Sammon's map showing topological relations between 3,097 discriminative patterns. (B) Dendrogram showing results of agglomerative clustering of 3,097 discriminative patterns into 4 super-patterns. (C) Geographical distribution of footprints of the 4identified super-patterns. (D) Meaning of 4 super-patterns in terms of socio-economic indicators.