ESTATE: Strategy for Exploring Labeled Spatial Datasets Using Association Analysis

Tomasz F. Stepinski[1] Josue Salazar[1] Wei Ding[2] and Denis White[3]

[1] Lunar and Planetary Institute, Houston, TX 77058, USA
tom@lpi.usra.edu salazar@lpi.usra.edu
[2] Department of Computer Science, University of Massachusetts Boston, Boston, MA 02125, USA
ding@cs.umb.edu
[3] US Environmental Protection Agency, Corvallis, OR 97333, USA
white.denis@epa.gov

Abstract. We propose an association analysis-based strategy for exploration of multi-attribute spatial datasets possessing naturally arising classification. Proposed strategy, ESTATE (**E**xploring **S**patial da**T**a **A**ssociation pat**TE**rns), inverts such classification by interpreting different classes found in the dataset in terms of sets of discriminative patterns of its attributes. It consists of several core steps 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. Detailed description of the ESTATE framework is followed by its application to the domain of ecology using a dataset that fuses the information on geographical distribution of biodiversity of bird species across the contiguous United States with distributions of 32 environmental variables across the same area.

Key words: spatial databases, association patterns, clustering, similarity measure, biodiversity

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 of 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. The need for such classification-in-reverse tool arises in many domains, including cases that may influence economic and political decisions and have significant societal repercussions. For example, a fusion of election results with socio-economic indicators form an administrative region-based spatial dataset that can be explored using ESTATE to reveal a spatio-socio-economic makeup of electoral support for different office seekers [23]. The framework can be also utilized for analyzing a diversity of underlying drivers of change (temporal, spatial, and modal) in the spatial system. An expository example of spatial change analysis – pertaining to geographical distribution of biodiversity of bird species across the contiguous United States – is presented in this paper.

The ESTATE interprets the divisions within the dataset by exploring the structure of the dataset. The strategy is underpinned by the framework of association analysis [1, 12, 34] that assures that complex interactions between all attributes are accounted for in a model-free fashion. Specifically, we rely on the contrast data mining [9, 2], 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.

The ESTATE framework consists of a number of independent modules; some of them are based on existing techniques while others represent original contributions. We present two original contributions to the field of data mining: 1) a novel similarity measure between itemsets that makes possible clustering of transactional patterns thus enabling effective summarization of thousands of discovered nuggets of knowledge, and 2) a strategy for disambiguation of class labels in datasets where classification is not naturally present and needs to be deduced from the character of the dataset.

2 Related Work

There is a vast literature devoted to classification/prediction techniques. In the context of spatial (especially, geospatial) datasets many broadly used predictors are based on the principle of regression, including multiple regression [25], logistic regression [30, 7, 14], Geographically Weighted Regression (GWR) [4, 11], and kernel logistic regression [29]. These techniques are ill-suited for our

stated purpose. A machine learning-based classifier could be constructed for the dataset where all objects have prior labels (usually, there is no need to do it). Denoting a classifier function as $F : F(attributes) \rightarrow class$, its inverse $F^{-1} : F^{-1}(class) \rightarrow attributes$ would give a set of all of the objects (their attribute vectors) mapped to a given class. However, the outcome of F^{-1} would be of no help to our purpose because it does not provide any synthesis leading to the understanding the common characteristics of the objects belonging to a given class. The exception is the classification and regression tree (CART) classifier, whose hierarchical form of F allows interpretation of F^{-1}. Indeed, the use of regression trees was proposed [28] to map spatial divisions of class variable. Our association analysis-based approach provides a more natural, data-centric alternative approach to the regression trees. The possibility of using transactional patterns for exploration of spatial datasets received little attention. Application of association analysis to geospatial data was discussed in [10, 22], and another application, to the land cover change was discussed in [19]. These studies did not utilized discriminative pattern mining. In addition, they lack any pattern synthesis techniques making the results difficult to interpret by domain scientists.

One of the major challenges of association analysis is the explosive number of identified patterns which leads to a need for pattern summarization. The two major approaches to pattern summarization are lossless and lossy representations. Lossless compression techniques include closed itemsets [20] and nonderivable itemsets [6]. In general, reduction in a number of patterns due to a lossless compression is insufficient to significantly improve interpretability of the results. More radical summarization is achieved via lossy compression techniques including maximal frequent patterns [3], top-k frequent patterns [13], top-k redundancy-aware patterns [26], profile patterns [32], δ-cover compressed patterns [31], and regression-based summarization [16].These techniques have been developed for categorical datasets where a notion of similarity between the items does not exist. The datasets we wish to explore with ESTATE are ordinal. We exploit the existence of an ordering information in the attributes of items to define a novel similarity between the itemsets. Our preliminary work on application of association analysis to exploration of spatial datasets is documented in [8, 24].

3 ESTATE Framework

The ESTATE framework is applied to a dataset composed of spatial objects characterized by their geographical coordinates, attributes, and class labels. The spatial dataset can be in the form of a raster (objects are individual pixels), point data (objects are individual points), or shapefile (objects are polygons). Information in each object is structured as follows $o = \{x, y; f_1, f_2, ..., f_m; c\}$, where x and y are object's spatial coordinates, f_i, $i = 1, \ldots, m$, are values of m attributes as measured at (x, y), and c is the class label. From the point of view of association analysis, each object (after disregarding its spatial coordinates and its class label) is a transaction containing a set of exactly m items $\{f_1, f_2, ..., f_m\}$,

which are assumed to have ordinal values. The entire spatial dataset can be viewed as a set of N fixed-length transactions, where N is the size of the dataset.

An itemset (hereafter also referred to as a pattern) is a set of items contained in a transaction. For example, assuming $m = 10$, $P = \{2, \_, \_, \_, 3, \_, \_, \_, \_, \_\}$ is a pattern indicating that $f_1 = 2$, $f_5 = 3$ while the values of all other attributes are not a part of this pattern. A transaction *supports* an itemset if the itemset is a subset of this transaction; the number of all transactions supporting a pattern is refereed to as a *support* of this pattern. For example, any transaction with $f_1 = 2$, $f_5 = 3$ "supports" pattern P regardless of the values of attributes in slots denoted by an underscore symbol in the representation of P given above. The support of pattern P is the number of transactions with $f_1 = 2$, $f_5 = 3$. Because transactions have spatial locations, there is also a spatial manifestation of support which we call a *footprint* of a pattern. For example a footprint of P is a set of spatial objects characterized by $f_1 = 2$, $f_5 = 3$.

The ESTATE framework consists of the following modules: (1) Mining for associative patterns that discriminate between two classes in the dataset (Section 3.1). (2) Disambiguating class labels so the divisions of objects into different classes coincide with footprints of discriminative patterns (Section 3.2). (3) Clustering all discriminative patterns into a small number of clusters representing diverse motifs of attributes associated with a contrast between the two classes (Section 3.3). (4) Visualizing the results in both attribute and spatial domains (see the case study in Section 4).

3.1 Mining for discriminative patterns

Without loss of generality we consider the case of the dataset with only two classes: $c = 1$ and $c = 0$. A *discriminating* pattern X is an itemset that has much larger support within a set of transactions $\mathcal{O}_\mathrm{p}$ stemming from $c = 1$ objects than within a set of transactions $\mathcal{O}_\mathrm{n}$ stemming from $c = 0$ objects. For a pattern X to be accepted as a discriminating pattern, its growth rate, $\frac{\sup(X,\mathcal{O}_\mathrm{p})}{\sup(X,\mathcal{O}_\mathrm{n})}$, must exceed a predefined threshold δ, where $sup(X, \mathcal{O})$ is the support of X in a dataset $\mathcal{O}$.

We mine for *closed* patterns that are *relatively* frequent in $\mathcal{O}^0_\mathrm{p}$. A pattern is frequent if its support (in $\mathcal{O}^0_\mathrm{p}$) is larger than a predefined threshold. Mining for frequent patterns reduces computational cost. Further significant reduction in computational cost is achieved by mining only for frequent closed patterns [21]. A closed pattern is a maximal set of items shared by a set of transactions. A closed pattern can be viewed as lossless compression of all non-closed patterns that can be derived from it. Mining only for closed patterns makes physical and computational sense inasmuch as closed patterns give the most detailed motifs of attributes associated with difference between the two classes.

3.2 Disambiguating class labels

In many (but not all) practical application, the class labels are implicit rather than explicit. For example, biodiversity index is continuously distributed across

the United States without a naturally occurring boundary between "high biodiversity" (class $c = 1$) and "not-high biodiversity" (class $c = 0$) objects. This introduces a question of what is the best way to partition the dataset into the two classes? One way is to divide the objects using distribution-deduced threshold on the class variable, another is to use the union of footprints of mined discriminative patterns. These two methods will result in different partitions of the dataset introducing potential ambiguity to class labels. We propose to disambiguate the labeling by iterating between the two definitions until the two partitions are as close to each other as possible.

We first calculate the initial $\mathcal{O}_{\mathrm{p}}^0$–$\mathcal{O}_{\mathrm{n}}^0$ partition using a threshold on the value of the class variable. Using this initial partition, our algorithm mines for discriminating patterns. We calculate a footprint of each pattern and the union of all footprints. The union of the footprints intersects, but is not identical to the footprint of $\mathcal{O}_{\mathrm{p}}^0$. Second, we calculate the next iteration of the partition $\mathcal{O}_{\mathrm{p}}^1$–$\mathcal{O}_{\mathrm{n}}^1$ and the new set of discriminating patterns. The objects that were initially in $\mathcal{O}_{\mathrm{n}}^0$ are added to $\mathcal{O}_{\mathrm{p}}^1$ if they are in the union of footprints of the patterns calculated in first step, their values of class variable are "high enough", and they are neighbors of $\mathcal{O}_{\mathrm{p}}^0$. Because of this last requirement, the second step is in itself an iterative procedure. The requirement that incorporated objects have "high enough" values of class variable is fulfilled by defining a buffer zone. The buffer zone is easily defined in a dataset of ordinal values; it consists of objects having a value one less than the minimum value allowed in $\mathcal{O}_{\mathrm{p}}^0$. Finally, we repeat the second step calculating $\mathcal{O}_{\mathrm{p}}^{\mathrm{i}}$ and its corresponding set of discriminating patterns from the results of $\mathrm{i}-1$ iteration until the iteration process converges. Note that convergence is assured by the design of the process. The result is the optimal $\mathcal{O}_{\mathrm{p}}$–$\mathcal{O}_{\mathrm{n}}$ partition and the optimal set of discriminating patterns.

3.3 Pattern Similarity Measure

Despite considering only frequent closed discriminative patterns, the ESTATE finds thousands of patterns. A single pattern provides a specific combination of attribute values found in a specific subset of the $c = 1$ class of objects but nonexistent or rare among $c = 0$ class objects. The more specific (longer) the pattern the smaller is its footprint; patterns having larger spatial presence tend to be less specific (shorter). Because of this tradeoff there is not much we can learn about the global structure of the dataset from a single pattern; such pattern provides either little information on regional scale or a lot of information on local scale. In order to effectively explore the entire dataset we need to consider all mined patterns each covering only relatively small spatial patch, but together covering the entire domain of the $c = 1$ class. To enable such exploration we cluster the patterns into larger aggregates of similar patterns by taking advantage of ordering information contained in ordinal attributes of spatial objects. The clustering is made possible by the introduction of a similarity measure between the patterns. We propose to measure a similarity between two patterns as a similarity between their footprints. Hereafter we will continue to refer to the

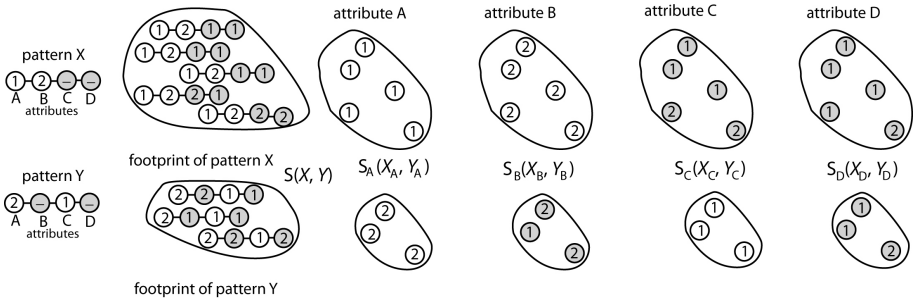

Fig. 1. Graphics illustrating the concept of similarity between two patterns. White items are part of the pattern, gray items are not the part of the pattern.

"pattern similarity measure" with the understanding that the term "pattern" is used as a shortcut for the set of objects in its footprint.

Fig. 1 illustrates the proposed concept of pattern similarity. In this simple example each object has four attributes denoted by A, B, C, and D, respectively. Each attribute has only one of two possible values: 1 or 2. Pattern $X = \{1, 2, \_, \_\}$ is supported by 5 objects and pattern $Y = \{2, \_, 1, \_\}$ is supported by 3 objects. The similarity between patterns X and Y is the similarity between the two sets of 4-dimensional vectors constructed from the values of items in transactions belonging to respective footprints. Similarity of each dimension (attribute) is calculated separately as a similarity between two sets of scalar entities. The total similarity is the weighted sum of the similarities of all attributes.

The similarity between patterns X and Y is $S(X,Y) = \sum_{i=1}^{m} w_i S_i(X_i, Y_i)$, where X_i, Y_i indicate the ith attribute, w_i indicates the ith weight (we use $w_i = 1$ in our calculations), and m is the number of attributes. The similarity between ith attribute in the two patterns $S_i(X_i, Y_i)$ is calculated using group average, a technique similar to the UPGMA (Unweighted Pair Group Method with Arithmetic mean) [18] method of calculating linkage in agglomerative clustering. The UPGMA method reduces to $S_i(X_i, Y_i) = s(x_i, y_i)$ for attributes which are present in both patterns (like an attribute A in an example shown in Fig. 1); here x_i and y_i are the values of attributes X_i and Y_i ($x_A = 1$ and $y_A = 2$ in the example on Fig. 1) and $s(x_i, y_i)$ is the similarity between those values (see below). If the ith attribute is present in the pattern Y but absent in the pattern X (like an attribute C in an example shown in Fig. 1) the UPGMA method reduces to

$$S(-, Y_i) = \sum_{k=1}^{n} P_X(x_k) s(z_k, y_i) \tag{1}$$

where $P_X(x_k)$ is the probability of ith attribute having the value x_k in all objects belonging to the footprint of X and n is the number of different values the ith attribute can have. The UPGMA reduces to an analogous formula if the ith attribute is present in the pattern X but it's absent in the pattern Y (like an attribute B in an example shown in Fig. 1). Finally, if the ith attribute is absent in both patterns (like an attribute D in an example shown in Fig. 1) the UPGMA

gives

$$S(-_i, -_i) = \sum_{l=1}^{n} \sum_{k=1}^{n} P_X(x_l) P_Y(y_k) s(x_l, y_k) \tag{2}$$

We propose to calculate the similarity between the two values of ith attribute using a measure inspired by an earlier concept of measuring similarities between ordinal variables using information theory [17]. The similarity between two ordinal values of same attribute $s(x_i, y_i)$ is measured by the ratio between the amount of information needed to state the commonality between x_i and y_i, and the information needed to fully describe both x_i and y_i.

$$s(x_i, y_i) = \frac{2 \times \log\ P(x_i \vee z_1 \vee z_2 \ldots \vee z_k \vee y_i)}{\log\ P(x_i) + \log\ P(y_i)} \tag{3}$$

where $z_1, z_2, \ldots, z_k$ are ordinal values such that $z_1 = x_i + 1$ and $z_k = y_i - 1$. Probabilities, $P()$, are calculated using the known distribution of the values of ith attribute in $\mathcal{O}_\mathrm{p}$.

Using a measure of "distance" ($dist(X, Y) = \frac{1}{S(X,Y)} - 1$) between each pair of patterns in the set of discriminative patterns we construct a distance matrix. In order to gain insight into the structure of the set of discriminative patterns we visualize the distance matrix using clustering heat map. The heat map is the distance matrix with its columns and rows rearranged to place rows and columns representing similar patterns near each other. We determine an appropriate order of rows and columns in the heat map by performing a hierarchical clustering (using an average linkage) of the set of discriminative patterns and sorting the rows and columns by the resultant dendrogram. The values of distances in the heat map are coded by a color gradient enabling the analyst to visually identify interesting clusters of patterns.

4 Case study: biodiversity of bird species

We apply the ESTATE framework to the case study pertaining to the discovery of associations between environmental factors and the spatial distribution of biodiversity across the contiguous United States. Roughly, biodiversity is a number of different species (of plants and/or animals) within a spatial region. A pressing problem in biodiversity studies is to find the optimal strategy for protecting the species given limited resources. In order to design such a strategy it is necessary to understand associations between environmental factors and the spatial distribution of biodiversity. In this context we apply ESTATE to discover existence of different environments (patterns or motifs of environmental factors) which associate with the high levels of biodiversity.

The database is composed of spatial accounting units resulting from tessellation of the US territory into equal area hexagons with center-to-center spacing of approximately 27 km. For each unit the measure of biodiversity (class variable) and the values of environmental variables (attributes) are given. The biodiversity measure is provided [33] by the number of species of birds exceeding a specific

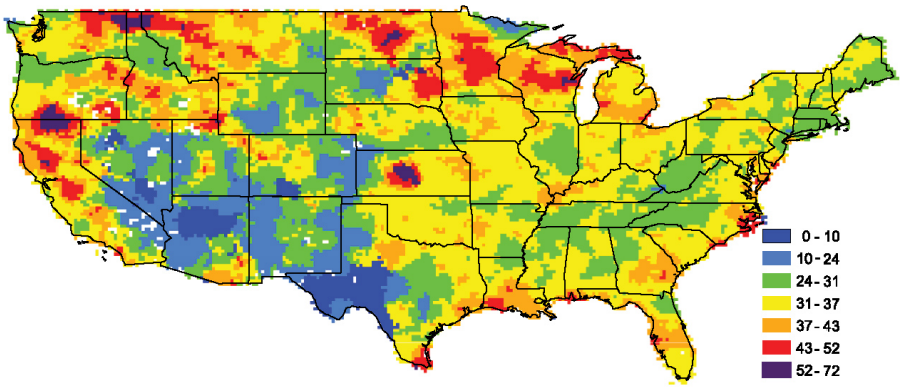

Fig. 2. Biodiversity of bird species across the contiguous United States. Two categories with the highest values of biodiversity (purple and red) are chosen as the initial high biodiversity region. Missing data regions are shown in white.

threshold of probability of occurrence in a given unit. Fig. 2 shows the distribution of biodiversity measure across the contiguous US. The environmental attributes [27] include terrain, climatic, landscape metric, land cover, and environmental stress variables that are hypothesized to influence biodiversity; we consider m=32 such attributes. The class variable and the attributes are discretized into up to seven ordinal categories (lowest, low, medium-low, medium, medium-high, high, highest) using the "natural breakes" method [15].

Because of the technical demands of the ESTATE label disambiguation module we have transformed the hexagon-based dataset into the square-based dataset. Each square unit (pixel) has a size of 22 $\times$ 22 km and there are N=21039 data-carrying pixels in the transformed dataset. The dataset does not have explicit labels. Because we are interested in contrasting the region characterized by high biodiversity with the region characterized by not-high biodiversity we have partitioned the dataset into $\mathcal{O}_{\rm p}$ corresponding to $c = 1$ class and consisting initially of the objects having high and highest categories of biodiversity and $\mathcal{O}_{\rm n}$ corresponding to $c = 0$ class and consisting initially of the objects having lowest to medium-high categories of biodiversity. The label disambiguation module modifies the initial partition during the consecutive rounds of discriminative data mining.

We identify frequent closed patterns discriminating between $\mathcal{O}_{\rm p}$ and $\mathcal{O}_{\rm n}$ using an efficient depth-first search method [5]. We mine for patterns having growth rate $\geq$50 and are fulfilled by at least 2% of transactions (pixels) in $\mathcal{O}_{\rm p}$. We also keep only the patterns that consist of eight or more attributes; shorter patterns are not specific enough to be of interest to us. We have found 1503 such patterns. The patterns have lengths between 8 and 20 attributes; the pattern length is broadly distributed with the maximum occurring at 12 attributes. Pattern size

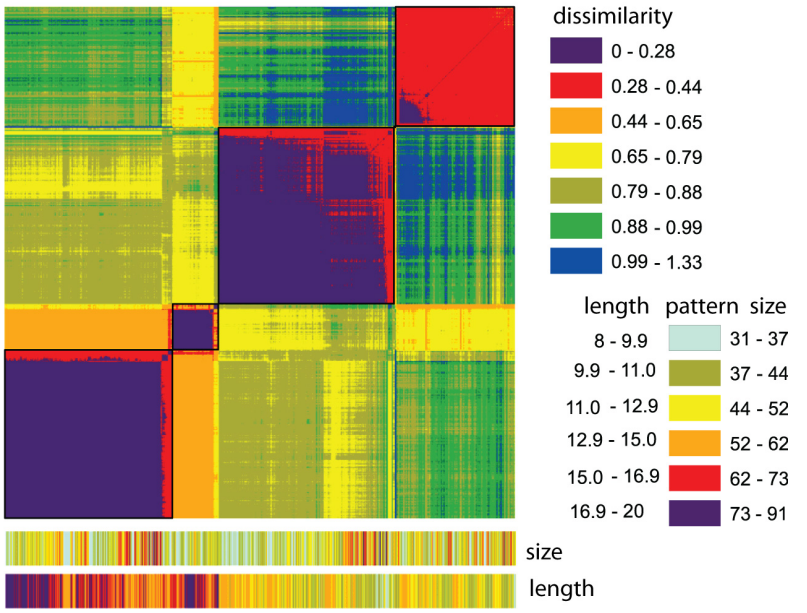

Fig. 3. Clustering heat map illustrating pairwise similarities between pairs of patterns in the set of 1503 discriminating patterns. The two bars below the heat map illustrate size of the pattern size and its length, respectively.

(support) varies from 31 to 91 pixels; the distribution of pattern size is skewed toward the high values and the maximum occurs at 40 pixels.

Fig. 3 shows a heat map constructed from a distance (dissimilarity) matrix calculated for all pairs of patterns in the set of 1503 patterns that discriminate between $\mathcal{O}_{\mathrm{p}}$ and $\mathcal{O}_{\mathrm{n}}$. The heat map is symmetric because distance between any two patterns is calculated twice. Deep purple and red colors indicate similar patterns whereas blue and green colors indicate dissimilar patterns. The heat map clearly shows that the entire set of discriminative patterns naturally breakes into four clusters as indicated by purple and red color blocks on the map. Indeed, there are five top level clusters, but the fourth cluster, counting from the lower left corner, has only 4 patterns and is not visible in the heat map at the scale of Fig. 3. The patterns in each cluster identify similar combinations (motifs) of environmental attributes that are associated with the region of high biodiversity. The visual analysis of the heat map indicates that there are four (five if we count the small 4-pattern cluster) distinct motifs of environmental attributes associated with high levels of biodiversity. Potentially, these motifs indicate existence of multiple environmental regimes that differ from each other but are all conducive to high levels of biodiversity.

The clusters can be characterized and compared from two different perspectives. First, we can synthesis the information contained in all patterns belonging to each cluster; this will yield combinations of attributes that set apart the region

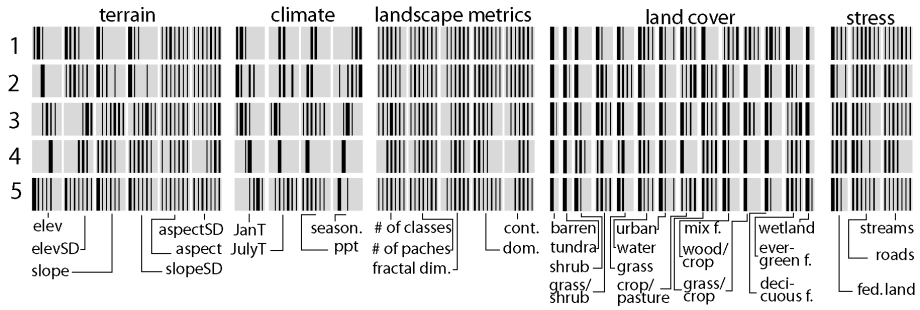

Fig. 4. Bar-code representation of the five regimes (clusters) of high biodiversity. See description in the main text.

associated with a given cluster from the not-high biodiversity region. Second, we can synthesize the information about prevailing attributes in the region associated with a given cluster; this will reveal a set of predominant environmental conditions associated with a given high biodiversity region (represented by a cluster). Because clusters are agglomerates of patterns and regions are agglomerates of transactions, they can be synthesized by their respective compositions. The biodiversity dataset has $m = 32$ attributes, thus each cluster (region) can be synthesized by 32 histograms, each corresponding to a composition of a particular attribute within a cluster (region). Our challenge is to present this large volume of information in a manner that is compact enough to facilitate immediate comparison between different clusters.

In this paper we restrict ourself to synthesizing and presenting the predominate environmental conditions associated with each of the five clusters identified in the heat map. Recall that the attributes are categorized into 7, 4, or 2 ordinal categories, thus a histogram representing a distribution of the values taken by an attribute in a given cluster consists of up to seven percentage-showing numbers. Altogether, 173 numbers, ranging in values from 0 (absence of a given attribute from cluster composition) to 1 (only a single value of a given attribute is present in a cluster) represents a summary of a cluster. We propose a bar-code representation of such summary. Such representation facilitates quick qualitative comparison between different clusters. Fig. 4 shows the bar-coded description for the five clusters corresponding to different biodiversity regimes. A cluster bar-code contains 32 fragments each describing a composition of a single attribute within a cluster. In Fig. 4 these fragments are grouped into five thematic categories: terrain (6 attributes), climate (4 attributes), landscape elements (5 attributes), land cover (14 attributes), and stress (3 attributes). Each fragment has up to seven vertical bars representing ordered categories of the attribute its represent. 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.

The five regimes of high biodiversity differs on the first four terrain attributes and all climate attributes. The landscape metrics attributes are similar except for

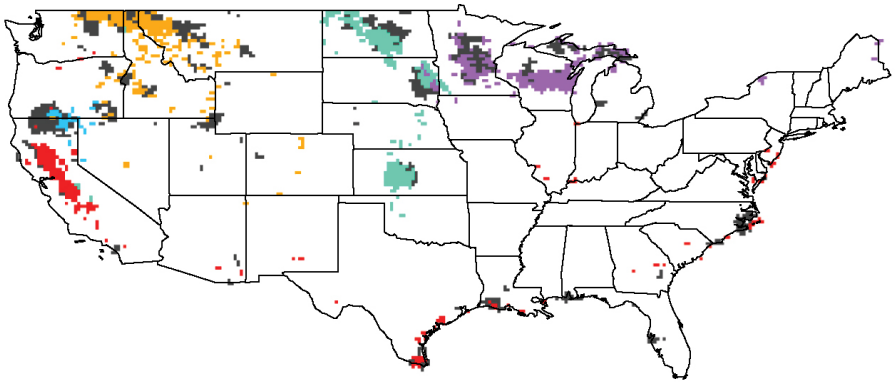

Fig. 5. Spatial footprints of five pattern clusters. White – not high biodiversity region; gray – high biodiversity region; purple (cluster #1), light green (cluster #2), yellow (clister #3), blue (cluster #4), and red (cluster #5) – footprints of the five clusters.

regime #4. Many land cover attributes are similar indicating that a number of land cover types, such as, for example, tundra, barren land or urban are absent in all high biodiversity regimes. More in depth investigation of the bar codes reveals that the regime #1 is dominated by the crop/pasture cover, the regime #2 by the wood/crop cover, the regime #3 by the evergreen forest, and the regimes # 4 and #5 are not dominated by any particular land cover. Finally, environmental stress attributes are similar except for the federal land that is more abundant within the regions defined by the regimes #3 and #4.

Spatial manifestation of the five clusters identified in the heat map are shown in Fig. 5 where transactions (pixels) fulfilled by patterns belonging to different clusters are indicated by different colors. Interestingly, different environmental regimes (clusters) are located at distinct geographical locations. This geographical separation of the clusters is the result and not a build-in feature of our method. In principle, footprints of different discriminative patterns may overlap, and footprints of the entire clusters may overlap as well. It is a property of the biodiversity dataset that clusters of similar discriminative patterns have non-overlapping footprints.

Note that in our calculations the label disambiguation module did not achieve complete reconciliation between the region of high biodiversity and the union of support of all discriminating patterns. The gray pixels on Fig. 4 indicate transactions that are in the $\mathcal{O}_{\mathrm{p}}$ but are not in the union of support of all the patterns. The ESTATE guarantees convergence of the disambiguation module but does not guarantee the complete reconciliation of the two regions. However, perfect correspondence is not required and, in fact, less than perfect correspondence provides some additional information. The gray areas on Fig. 4 represents atypical regions characterized by infrequent combinations of environmental attributes.

5 Discussion

A machine learning task of predicting labels of class variable using explanatory variables became an integral component of spatial analysis and is broadly utilized in many domains including geography, economy, and ecology. However, many interesting spatial datasets possesses natural labels, or their labels can be easily classified without resorting to machine-learning methods. We have developed the ESTATE framework in order to understand such naturally occurring divisions in terms of dataset attributes. In a broad sense, the purpose of ESTATE is reverse to the purpose of a classification.

Many real life problems analyzable by ESTATE may be formulated in terms of "spatial change" datasets (class labels change from one location to another). Other real life problems, analyzable by ESTATE, may be formulated as "temporal change" datasets (class labels indicate presence or absence of change in measurements taken at different times), or "modal change" datasets (class labels indicate agreement or disagreement between modeled and actual spatial system). An expository example given in Section 4 belongs to the spatial change dataset type. The biodiversity dataset has "natural" classes inasmuch as it can be divided into high and no-high biodiversity parts just on the basis of the distribution of biodiversity measure. Note that classes other than "high" can be as easily defined; for example, for a complete evaluation of the biodiversity dataset we would also define a "low" class. Other datasets (see, for example, [23]) have prior classes and require no additional pre-processing.

It is noted that ESTATE (like most other data discovery techniques) discovers associations and not causal relations. In the context of the biodiversity dataset it means that ESTATE has found five different environments that associate with high biodiversity but it does not proof actual causality between those environments and high levels of biodiversity. It is up to the domain experts to review the results and draw the conclusions. The causality is strongly suggested if the experts believe that the 32 attributes used in the calculation exhaust the set of viable controlling factors of biodiversity.

A crucial component of the ESTATE is the pattern similarity measure that enables clustering of similar patterns into agglomerates. We stress that our method does not use patterns to cluster objects, instead patterns themselves (more precisely their footprints) are the subject of clustering. This methodology can be applied outside of the ESTATE framework for summarization of any transactional patterns as long as their items consist of ordinal variables. Future research would address how to extend our similarity measure to categorical variables.

Acknowledgements

This work was partially supported by the National Science Foundation under Grant IIS-0812271.

References

1. R. Agrawal and A. N. Swami. Fast algorithms for mining association rules. In *Proceedings of VLDB*, page 487499, 1994.
2. S. D. Bay and M. J. Pazzani. Detecting change in categorical data: Mining contrast sets. In *Knowledge Discovery and Data Mining*, pages 302–306, 1999.
3. R. J. Bayardo, Jr. Efficiently mining long patterns from databases. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 85–93, Seattle, Washington, United States, 1998.
4. C. A. Brunsdon, A. S. Fotheringham, and M. B. Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28:281–298, 1996.
5. D. Burdick, M. Calimlim, and J. Gehrke. MAFIA: a maximal frequent itemset algorithm for transactional databases. In *Proceedings of the 17th international conference on data engineering. Heidelberg, Germany*, 2001.
6. T. Calders and B. Goethals. Non-derivable itemset mining. *Data Min. Knowl. Discov.*, 14(1):171–206, 2007.
7. J. Cheng and I. Masser. Urban growth pattern modeling: a case study of wuhan city, pr china. *Landscape and Urban Planning*, 62(4):199–217, 2003.
8. W. Ding, T. F. Stepinski, and J. Salazar. Discovery of geospatial discriminating patterns from remote sensing datasets. In *Proceedings of SIAM International Conference on Data Mining*, 2009.
9. G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52, San Diego, California, United States, 1999.
10. J. Dong, W. Perrizo, Q. Ding, and J. Zhou. The application of association rule mining to remotely sensed data. In . 345, editor, *Proc. of the 2000 ACM symposium on Applied computing*, 2000.
11. A. S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically Weighted Regression: the analysis of spatially varying relationships*. Chichester: Wiley, 2002.
12. J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1):5387, 2004.
13. J. Han, J. Wang, Y. Lu, and P. Tzvetkov. Mining top k frequent closed patterns without minimum support. In *ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining*, page 211, Washington, DC, USA, 2002.
14. Z. Hu and C. Lo. Modeling urban growth in atlanta using logistic regression. *Computers, Environment and Urban Systems*, 31(6):667–688, 2007.
15. G. F. Jenks. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7:186–190, 1967.
16. R. Jin, M. Abu-Ata, Y. Xiang, and N. Ruan. Effective and efficient itemset pattern summarization: regression-based approaches. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 399–407, Las Vegas, Nevada, USA, 2008.
17. D. Lin. An information-theoretic definition of similarity. In *International Conference on Machine Learning*, Madison, Wisconsin, July 1998.
18. L. McQuitty. Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, 26:825–831, 1966.

19. J. Mennis and J. W. Liu. Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Transactions in GIS*, 9(1):5–17, 2005.
20. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *ICDT'99: Proceeding of the 7th International Conference on Database Theory*, pages 398–416, 1999.
21. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *ICDT '99: Proceedings of the 7th International Conference on Database Theory*, pages 398–416, 1999.
22. U. Rajasekar and Q. Weng. Application of association rule mining for exploring the relationship between urban land surface temperature and biophysical/social parameters. *Photogrammetric Engineering & Remote Sensing*, 75(3):385–396, 2009.
23. T. Stepinski, J. Salazar, and W. Ding. Discovering spatio-social motifs of electoral support using discriminative pattern mining. In *proceedings of COM.Geo 2010 1st International Conference on Computing for Geospatial Reserch & Applications*, 2010.
24. T. F. Stepinski, W. Ding, and C. F. Eick. Controlling patterns of geospatial phenomena. *submitted to Geoinformatica*, 2009.
25. D. M. Theobald and N. T. Hobbs. Forecasting rural land use change: a comparison of regression and spatial transition-based models. *Geographical and Environmental Modeling*, 2:65–82, 1998.
26. C. Wang and S. Parthasarathy. Summarizing itemset patterns using probabilistic models. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 730–735, Philadelphia, PA, USA, 2006.
27. D. White, B. Preston, K. Freemark, and A. Kiester. A hierarchical framework for conserving biodiversity. In J. Klopatek and R. Gardner, editors, *Landscape ecological analysis: issues and applications*, pages 127–153. New York: Springer-Verlag, 1999.
28. D. White and J. C. Sifneos. Regression tree cartography. *J. Computational and Graphical Statistics*, 11 (3):600–614, 2002.
29. B. Wu, B. Huang, and T. Fung. Projection of land use change patterns using kernel logistic regression. *Photogrammetric Engineering & Remote Sensing*, 75(8):971–979, 2009.
30. F. Wu and A. G. Yeh. Changing spatial distribution and determinants of land development in chinese cities in the transition from a centrally planned economy to a socialist market economy: A case study of guangzhou. *Urban Studies*, 34(11):1851–1879, 1997.
31. D. Xin, J. Han, X. Yan, and H. Cheng. Mining compressed frequent-pattern sets. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 709–720, Trondheim, Norway, 2005.
32. X. Yan, H. Cheng, J. Han, and D. Xin. Summarizing itemset patterns: a profile-based approach. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 314–323, Chicago, Illinois, USA, 2005.
33. K. Yang, D. Carr, and R. O'Connor. Smoothing of breeding bird survey data to produce national biodiversity estimates. In *Computing Science and Statistics, Proceeding of the 27th Symposium on the Interface*, pages 405–409, 1995.
34. M. Zaki and M. Ogihara. Theoretical foundations of association rules. In *the 3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1998.