Learning Approximate Thematic Maps from Labeled Geospatial Data*

(Extended Abstract)

Mehdi Sharifzadeh, Cyrus Shahabi, Craig A. Knoblock
Computer Science Department and Information Sciences Institute
University of Southern California
Los Angeles, California 90089

[sharifza, shahabi]@usc.edu, knoblock@isi.edu

ABSTRACT

Building accurate thematic maps which show distribution of a feature over a geographic area is a challenging task when the sample dataset is limited in size and distribution. Classification of these geospatial datasets is a promising approach towards building approximate thematic maps. Moreover, choosing the appropriate classification method that considers spatial autocorrelation in data would result into more accurate maps. This paper investigates the application of different classification methods on real spatial datasets. We study how factors such as distribution of the training data, neighborhood relationships and geometry of the original map can affect the accuracy of the generated map. Consequently, we report on measurements comparing the accuracy of the investigated methods on different datasets. Our experimental setup benefits from a spatial database system to compare the regions of the approximate map with those of the original accurate map.

1. INTRODUCTION

Recent developments in the data collection techniques including remote sensing and sensor networks as well as geocoded customer addresses in transactional systems have resulted into the storage of huge amounts of geospatial *objects* in databases. The research area of spatial data mining utilizes algorithms and techniques from statistics, machine learning, spatial reasoning and spatial databases to realize various spatial relationships among these data objects. Spatial classification is one of these techniques that analyzes spatial and non-spatial attributes of the data objects to partition the data into a set of classes. These classes generate a map representing groups of related data objects. To illustrate, data objects can be houses each with spatial *geocoordinate* and non-spatial zip code attributes. Spatial classification of the geocoordinates based on the objects' zip code values (i.e., *features*) would generate an approximate *thematic map* of the zip code areas. Although there have been some research studies on classifying spatial datasets [4], almost no method has used the visual representation of the results in order to evaluate their accuracy.

Maps have been extensively used as the main references in the field of geography. They are the most common tools for visualizing geospatial datasets. In addition, thematic maps show the distribution of a feature over a limited geographic area. They illustrate how an area can be divided into different labeled regions. In most of the cases, these maps are approximated using a limited set of labeled data points located inside the desired area. As another example, with the sensor network application domain, suppose thousands of sensors with GPS systems are deployed in a battle field monitoring the chemicals in the air. One may be interested to build the approximate thematic map for the density level of chemicals in the air from the data monitored by the sensors. In this paper, we use different classification methods to approximately generate thematic maps.

We study the application of four classification methods and evaluate the accuracy of each of these approaches using its traditional test approach. We identify how factors such as distribution of the training data, neighborhood relationships and geometry of the original map can affect the accuracy of the approximate map. Our experimental results verify that classifying uniform test datasets is not enough to evaluate the accuracy of an approximate map. We propose to use more accurate measure values that compare the geometry of the original and approximate maps. Using features of a spatial database we define area-based precision and recall values that compare the areas of each region in the approximate map and its corresponding region in the original map. We plan to continue our evaluation for other spatial classification methods.

2. DEFINITIONS

We first define the main terms used throughout the paper and describe their characteristics. Then we formally describe the problem and discuss how it is related to the classification problem domain.

2.1 Problem Components

Each data object in our application domain is a 2-dimensional **point** in geographic space, in the form of (*Longitude, Latitude*). These coordinates can be generated from a valid street address using a geocoder. Although a location is a larger entity defined as a set of neighboring points, we will use the point and location interchangeably.

Any non-spatial attribute of a location is called a *theme* or a **feature**. Two different types of features exist. A class of features such as *zip code* or *phone area code* is assigned to every single location in geographic space. Thus, each location is *labeled* with a feature value. A different class of

*This research has been funded in part by NSF grants EEC-9529152 (IMSC ERC), IIS-0082826 (ITR), IIS-0238560 (CAREER) and IIS-0302168 (ITR), DARPA and USAF under agreement nr. F30602-99-1-0524, and unrestricted cash gifts from Okawa Foundation and Microsoft.

features such as *population* are maintained for larger areas. The value of these features has no meaning/use when defined for a specific location. With our examples, zip codes and MSA codes (see Section 4) are two different features and different values of each feature corresponds to different class labels. We will refer to class labels and feature values as features.

Thematic Map is a map primarily designed to show a theme, a single spatial distribution or a pattern, using a specific map type [2]. These maps show the distribution of a feature over a limited geographic area. Each map defines a partitioning of the area into a set of closed and disjoint regions, each includes all the points with the same feature value. Formally speaking, a thematic map is a partitioning of 2-d space into *disjoint regions* P_i, $(i = 1, 2, \ldots, m)$ so that:

1. Each partition region P_i is corresponding to one feature value $F(P_i)$ but one feature value can be assigned to several regions. Therefore there is a one-to-many mapping from feature space to region space. In this paper, we focus on the maps with a one-to-one mapping between regions and features.

2. For each point o inside region P_i, the feature value of o is equivalent to that of P_i (i.e., $F(o) = F(P_i)$).

2.2 Problem Definition

Official organizations usually define thematic maps with strictly defined boundaries. For example, U.S. Postal Service specifies zip code maps for each state in the United States. We call each of these accurate maps an *original map*. Consider a situation when such an original map is not available. However, a set of data points precisely labeled with the corresponding feature values is given. The problem is to find a method to create the best approximate map from the given sample points. In other words, we want to find a partitioning of 2-d space into disjoint regions P_i, $(i = 1, 2, \ldots, m)$ such that:

1. Each partition region P_i corresponds to one and only one feature value $F(P_i)$.

2. For each point o inside region P_i, $f \neq F(P_i)$:
$Probability(F(o) = F(P_i)) \quad > \quad Probability(F(o) = f)$

3. CLASSIFICATION METHODS

From a machine learning perspective, the thematic map problem is addressable using the spatial multi-class classification methods. That is, as the training points are geospatial coordinates in space, we should employ a classification algorithm which respects spatial relation between points (e.g., neighborhood information). The algorithm should generate decision boundaries for all feature classes in order to generate the desired map.

The task of classification is labeling a data object with a label from a given set of class labels based on the attributes of the object. Moreover, spatial classification benefits from the fact that closer points in the original space are more related to each other and hence more likely belong to the same class. Machine learning literature includes extensive research work on classification algorithms.

Characteristics of the training data and the corresponding accurate original map make us to be more selective in the approach we use. The data is accurate and the solution needs the most accurate region boundaries in the original space. Hence, the method must have a geometric interpretation in the point space. Motivated by the above requirements, we describe four different approaches and their application to generate the approximate map. In particular, we focus on *Nearest Neighbor*, *Linear and Quadratic Discriminant Analysis* and *Support Vector Machines* in turn.

3.1 The Nearest Neighbor Method

Tolber's first law of geography says "*everything is related to everything else, but nearby things are more related than distant things*" [6]. This fact implies *spatial autocorrelation* for the features in a geographic space. It means that there is a relation between features in neighboring points. This inspires us to use the *Nearest Neighbor* method for classifying point datasets. This method first stores all the training points with their labels. Subsequently, for any new point, it assigns the feature of the closest point in the training set to that point. Therefore, there is a unique feature assignment for each point.

The nearest neighbor algorithm does not explicitly compute decision boundaries for each feature. However, the decision boundaries form a subset of the Voronoi diagram for the training data. A Voronoi diagram [5] is the partitioning of a plane with n points into n convex polygons (Voronoi cells) such that each polygon contains exactly one point and every other point in a given polygon is closer to its central point than to any other point. Merging Voronoi cells corresponding to the points with the same feature value forms the map region for that value.

We implemented the nearest neighbor method by building Voronoi diagram for each set of data points. This approach enabled us to precisely compare the approximate map with the original map. First, an open source program, *qhull* [1], was used to generate Voronoi diagrams. As the next step, we find all adjacent Voronoi cells with the same feature and merge the areas they cover to produce the map region corresponding to that feature. A spatial database system, Informix Dynamic Server with Spatial Datablades [3], which provides spatial operations for handling geometry objects, was used for the merging step. Finally, we compared each region polygon to the corresponding region in the original map in order to measure precision-recall values.

We used approximate region as the retrieved set and the original region as the relevant set to define precision and recall values as follows:

$$Precision = \frac{Area(Intersection(approximate, original))}{Area(approximate)}$$

$$Recall = \frac{Area(Intersection(approximate, original))}{Area(original)}$$

We refer to the precision-recall values computed above as the *area-based precision-recall*. These measure values are different from traditional *test-based* values which are computed by counting the number of correctly classified data points in a test dataset.

3.2 Linear/Quadratic Discriminant Analysis

The main building blocks of a map are partition regions that are defined by their boundaries. Different discriminant functions try to approximately specify these *decision* boundaries. One interesting instance of such functions is a density estimator that relies on density of the points in each region.

Linear Discriminant Analysis (LDA) is a classification method which uses Gaussian density estimators as discriminant functions. LDA models each class density with a multivariate Gaussian and assigns a common covariance matrix to all classes. Quadratic Discriminant Analysis (QDA) is a generalization of LDA where each class can have different covariance matrices. Since LDA and QDA specify decision

boundaries between original data points without changing the shape and location of the data, we choose them as our next candidate methods for classifying the point data. We studied the impact of the training data density on our approximation results using these functions.

3.3 Support Vector Machines

Support Vector Machines (SVM) [7] are widely used in classifying large datasets. Different kernel functions incorporated into the main algorithm results into a flexible regression/classification tool. SVM maps all the training data points into a high-dimensional Hilbert space and then generates region boundaries as hyperplanes separating data points in that space. This training phase is expensive as SVM tries to solve a quadratic problem with as many variables as data points. This causes the original approach to be slow for large datasets. Therefore, researchers have proposed several optimized versions that we use in our experiments.

Original SVM algorithm provided by Vapnik [7] is a two-class learning method but there are some approaches to extend it to multi-class problems. SVM solves n class problems ($n > 2$) in two ways: 1) trains n machines, each classifying one class against the rest, 2) trains $n(n-1)/2$ machines, each classifying one class against one other class and uses a voting schema for each machine. We used the first approach in our experiments.

Since we expect the best possible trained SVM with the least error, we set the value of parameter C in SVM configuration to a large number. We globally scaled point attributes (latitude and longitude) as they were of the same domain type. Furthermore, to make the program train SVM using large training data, the chunking option was enabled. In our experiments, we trained SVM with four different kernels: radial basis, linear and polynomial kernels with the degrees of 2 and 3.

4. EXPERIMENTS

We conducted several studies to compare the precision of different classification methods and study the impact of the following factors on the accuracy of each approach: 1) density of the training data (point density), 2) distribution of the training data, and 3) complexity of the original map. The precision-recall values were used to measure how precisely each approach classifies different features in the result sets.

For our experiments, we used a real dataset for the United States obtained from U.S. Geological Survey (USGS). We extracted four different datasets from USGS data using different businesses in that area. We used zip code maps and U.S. Metropolitan Statistical Areas (MSA) as original maps in different experiments. The other dataset is the result of geocoding a set of valid addresses in the city of Los Angeles. We retrieved these addresses by querying an online White Pages service on Internet and defined zip code of each point as the feature. The key differences between these two different datasets are in the distribution and density of the points and the complexity of the original map. USGS data is uniformly distributed over the area with different densities for different businesses while white pages data is nonuniform and dense near the center of each feature region.

In one of the experiments, we investigated how precisely each classification method can approximate the original map. We used a different measure, *Harmonic Mean*, to combine area-based precision and recall values into a single value. Harmonic mean is defined as $2/(1/Recall + 1/Precision)$. It is equal to 1 when both precision and recall are 100% and 0 when one of them is close to 0. Figure 1 depicts the accuracy of the four different methods we used to approximate zip map for USGS data considering different densities for the training datasets. We made samples including different subsets of USGS data as our training datasets with different point densities. Then, we used each method to classify the training datasets and determine the corresponding boundaries. We computed the area-based accuracy measures by comparing generated map regions with the original zip map provided by USGS. As illustrated in the figure, as the point density in the training data grows, accuracy of all methods increases. SVM with a radial basis kernel shows the best accuracy even for the small point densities. Using linear and polynomial kernel of degree 2 in SVM method results the second and the third accurate maps. As QDA uses a quadratic kernel, it outperforms SVM with polynomial kernel of degree 3. Nearest neighbor and LDA methods create acceptable results with the precision up to 80% for dense training sets.

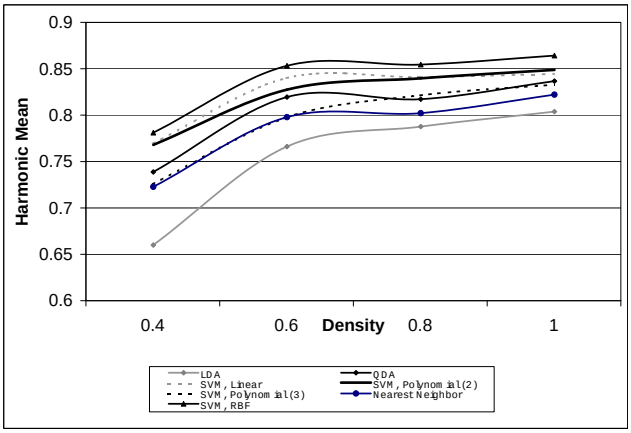

Figure 1: Harmonic Mean values for different methods generating zip map computed using area-based precision and recall values.

Throughout other experiments we studied the accuracy of the classification methods using area-based measures. Finally, the last set of experiments was aimed to study the impact of the training data distribution on the accuracy of the approximate map. The major results can be summarized as follows:

- Classification methods which generate decision boundaries for all classes can be applied to sample data points to build approximate thematic maps.
- Area-based precision values verify that SVM with a radial basis kernel outperforms all the other investigated methods.
- Area-based precision-recall values that provide more acceptable accuracy measures in practice, are usually smaller than their corresponding test-based values.
- A spatial database system can be efficiently used to compute area-based accuracy measures.
- Uniformly distributed features in the training dataset leads to a more accurate map for sparse datastes.
- The complexity of the original map indicates how precisely each classification method can build an approximate map.

5. CONCLUSION

We proposed to build approximate thematic maps using different classification methods. Through several empirical experiments we showed the accuracy of different methods using traditional test-based precision values. We introduced the area-based precision-recall, a more accurate measure, and performed a different set of experiments to compute these values using a spatial database system. We also studied the impact of the training dataset distribution on the generated approximate map.

6. REFERENCES

[1] Barber, C.B., Dobkin, D.P., and Huhdanpaa, H.T, *The Quickhull Algorithm for Convex Hulls* , ACM Transactions on Mathematical Software, 1996.

[2] Clarke, K.C., *Getting Started with GIS*, 2nd edition, 1999.

[3] Informix Corporation, *Informix Spatial Datablade Module*, Version 8.1, 2000.

[4] Koperski, K., Adhikary, J., Han J., Spatial Data Mining: Progress and Challenges, *Proceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, QB, 1996.

[5] Okabe, A., Boots, B., Sugihara, K., *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, John Wiley and Sons, Chichester, UK, 1992.

[6] Tobler, W.R., *Cellular Geography, Philosophy in Geography*, Gale and Olsson, Eds., Dordrecht, Reidel, 1979.

[7] Vapnik, V., *Statistical Learning Theory*, John Wiley and Sons, New York, 1998.