Learning Approximate Thematic Maps from Labeled Geospatial Data

M. Sharifzadeh, C. Shahabi & C. A. Knoblock
Computer Science Department and Information Sciences Institute
University of Southern California
Los Angeles, California 90089

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. We propose the classification of these geospatial datasets as a promising approach towards building approximate thematic maps. However, choosing an appropriate classification method that considers spatial autocorrelation in data is not trivial. This paper investigates the application of different classification methods on real-world 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 utilizes a spatial database system to compare the regions of the approximate map with those of the original accurate map. According to our experimental results, a Support Vector Machine (SVM) with a radial basis kernel outperforms all the other investigated methods.

1 INTRODUCTION

Recent developments in both data collection techniques through remote sensing and sensor networks and geocoding customer addresses in transactional systems have resulted in the availability of huge amounts of geospatial *objects* in databases. Moreover, the maturity of the spatial database technology which provides efficient storage and query capabilities for these bulky datasets has increased the opportunity of incorporating geospatial data into different application domains. Supporting spatial queries has been a promising step towards research on spatial data mining. 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 geospatial 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 number of classes. These classes can form a map representing various 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 studies on classifying spatial datasets (Koperski et al. 1996), to the best of our knowledge no study has used the visual representation of the results as a thematic map in order to evaluate the accuracy of its method. This is important when the main goal of the classification is to build thematic maps.

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 particular, 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 can be approximated using a limited set of labeled data points located inside the desired area. For example, in the domain of sensor networks, suppose thousands of sensors with GPS systems are deployed in a battle field

monitoring the chemicals in the air. One may be interested in building 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 various classification methods to generate approximate thematic maps. We study the application of four classification methods and evaluate the accuracy of each of these approaches using its traditional test procedure. The procedure evaluates how well the trained method can classify a test dataset and provides accuracy measures (*test-based* precision and recall). In addition, we propose to use more accurate measures that compare the *geometry* of the original and approximate maps. Using features of a spatial database we define our *area-based* precision and recall measures that compare the area of each region in the approximate map with its corresponding region in the original map. Finally, 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.

The remainder of the paper is organized as follows. Section 2 defines the main terms and characteristics of the problem. In Section 3, we describe some machine learning techniques used in classifying geospatial datasets. Section 4 includes our empirical experiments with real-world data and the results of applying different methods on labeled data objects. Section 5 reviews the geospatial interpolation techniques which are widely used in building thematic maps. Section 6 discusses the conclusions and our future plans.

2 DEFINITIONS

As the problem originates from the field of cartography and geography, we need to define some specific terms and identify their corresponding terms in the machine learning domain. 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 an extent 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 features such as *population* is maintained for extents. The value of these features has no meaning/use when defined for a specific point location. For our classification algorithms, zip codes and the US Metropolitan Statistical Areas (MSA) codes (see Section 4) are two different features whose different values correspond 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 (Clarke 2002). 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, ($i = 1, 2, \ldots, m$) such 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. $F(o) = F(P_i)$).

Figure 1 illustrates a California county map that can be viewed as a thematic map with county name as a feature. Throughout this paper, we will use map to refer to any thematic map.

Figure 1. California county map as a typical thematic map.

2.2 *Problem definition*

Official organizations usually define thematic maps with strictly defined boundaries. For example, US Postal Service specifies the zip code maps for each state in the United States. We call each of these accurate maps an *original map*. Consider the case 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, and feature $f \neq F(P_i)$:
   $Probability(F(o) = F(P_i)) \quad > \quad Probability(F(o) = f)$

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 exploits 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.

We should respect the characteristics of the training data and the corresponding accurate original map when choosing our classification approach. 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 discuss *Nearest Neighbor, Linear and Quadratic Discriminant Analysis* and *Support Vector Machines* in turn.

3.1 *The Nearest Neighbor method*

Tobler's first law of geography says "*everything is related to everything else, but nearby things are more related than distant things*" (Tobler 1979). 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 the new 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 (Okabe et al. 2000) 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. Figure 2 shows the way Voronoi diagrams can partition the space into map regions. Merging Voronoi cells corresponding to the points with identical features forms the map region for that value (more details is discussed in Section 4.2.1).

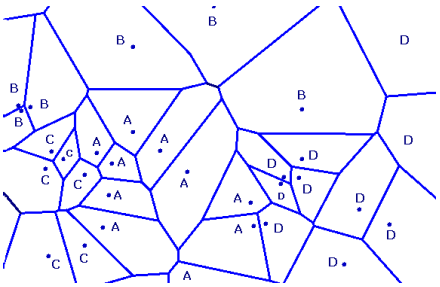

Figure 2. Voronoi diagram of a set of points with 4 different feature values.

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 in Section 4.

3.3 *Support Vector Machines*

Support Vector Machines (SVM) (Vapnik 1982; Vapnik 1998) are widely used in classifying large datasets. Different kernel functions incorporated into the main algorithm results in 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 an 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 is a two-class learning method but there are some approaches to extend it to multi-class problems. SVM can solve 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.

4 EXPERIMENTS

We conducted several experiments to compare the accuracy of different classification methods and study the impact of the following factors on the accuracy of each approach:

- d: density of the training data (point density). Our experiments were designed for different density levels in the training data.
- p: distribution of the training data. Uniform and nonuniform datasets were examined.
- c: complexity of the original map. We used two different original maps as our reference maps for measuring the accuracy.

The precision and recall measures were used to measure how precisely each approach classifies different features in the result sets. In the following sections, we describe different datasets and the way accuracy for each method was measured.

4.1 *Datasets*

For our experiments, we considered approximating two different original maps using two different datasets. We generated each map using both training datasets that included the data points labeled with the corresponding feature values. This combination results in four different experiments. As the original maps for these features are available, we can easily assign these labels to each data point by finding the map region which includes the point.

Our first dataset is a real-world dataset for the United States obtained from the US Geological Survey (USGS). The data uniformly covers a rectangular area with corner points latitude and longitude (21.25,-158.28) and (61.48,-67.94). Different businesses (e.g. schools and churches) in that area were used as data points in order to create an approximately uniform dataset. Using uniform sampling, we extracted four different datasets with different densities from the USGS data (density of the training points is defined as the number of points of interest over a one square mile area). Our second dataset is the result of geocoding a set of valid addresses in the city of Los Angeles. We retrieved these addresses by querying the data provided as an online White Pages service on the Internet (Verizon Inc. 2004). The addresses correspond to a set of restaurants located in an area of 30x30 miles. We used a geocoder application to convert these addresses to a set of 2-dimensional points in geographic space. We refer to the first dataset as USGS and the second one as WP.

The key difference between these two different datasets is in the distribution and density of the points. USGS data is uniformly distributed over the area with different densities for different businesses while WP data is nonuniform and dense near the center of each feature region.

Our first feature map consists of complicated regions of the US Metropolitan Statistical Areas (MSA). The US MSA represents geographic entities, defined by the United States Office of Management and Budget for use by the Federal statistical agencies, based on the concept of a core area with a large population nucleus, plus adjacent communities with a high degree of economic and social integration to that core. We used these maps as original maps and the MSA codes of the surrounding areas for WP and USGS points as their features. Figure 3 illustrates a small portion of these areas.

Our second map is the zip code map of the entire US. We used the zip code of each USGS and WP point as its feature for this map. As a result, we can precisely compare the approximate map generated by each approach with the original map.

Table 1 depicts the characteristics of our two original maps. It shows that the majority of the regions in the zip map are smaller and simpler than those of the US MSA map. Table 2 depicts characteristics of the two datasets we have used as our training data. It shows that there is a possibility that some data points in the dataset are labeled with no specific feature value. The classification methods generate a region for a certain feature value if and only if there is at least one point in the training dataset which is labeled with that value. As an example, any method which uses WP data to generate the zip map will generate an approximate map of only 203 regions out of 29,948 regions in the original map. In other words, the approximate map is a small portion of the original map as these 203 regions only cover a part of the city of Los Angeles.

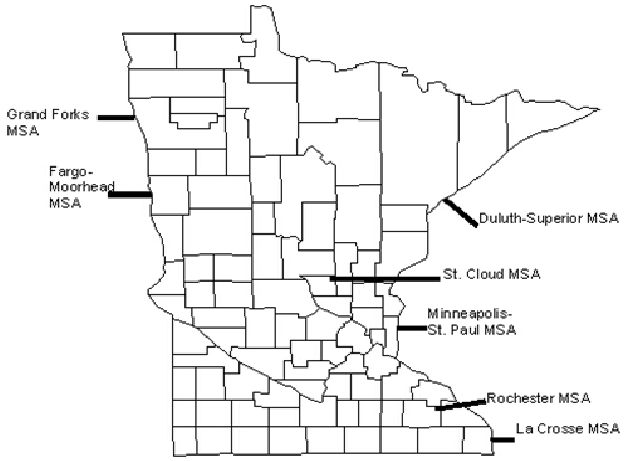

Figure 3. A subset of the US Metropolitan Statistical Areas in Minnesota.

Table 1. The zip map and the MSA map characteristics.

Parameter	MSA map	Zip map
Number of regions	314	29,948
Average area of regions (square mile)	2298.23	119.17
Total area covered (square mile)	721,645	3,568,836
Average number of vertices for each region	1585	70

4.2 *Implementations*

This section focuses on all database and mathematical tools we used to develop our experiments and compute the accuracy of each method.

4.2.1 *The Nearest Neighbor*

We implemented the nearest neighbor method by building the Voronoi diagram of each dataset. This approach enabled us to precisely compare the approximate map with the original map. First, an open source program, *qhull*, was used to generate the Voronoi diagrams (Barber et al. 1996). Next, we find all the adjacent Voronoi cells with an identical feature and merge their areas to produce the region corresponding to that feature. A spatial database system, Informix Dynamic Server featured with Spatial Datablades (Informix Corporation 2000), 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.

To illustrate, we show the above process through an example depicted in Figure 2. First, the Voronoi cell for each labeled point is created. We store each of these Voronoi cells, its corresponding point (Voronoi center) and the feature value itself as a tuple in the form of `(ST_Polygon, ST_Point, String)` in a relational table. Then, the following SQL statement returns polygons resulting from merging the Voronoi cells with the identical feature values:

```
SELECT Feature, ST_DISSOLVE(Voronoi_Cell)
FROM All_Voronoi_Cells
GROUP BY Feature;
```

Polygons retrieved by the above SQL statement form the approximate map generated by the nearest neighbor method. Figure 4 shows the merge step for one of the feature values (i.e. A).

Table 2. Various datasets used by our classification methods for training.

Dataset	Points	Size	Number of MSA values	Number of zip values
USGS	School	73,729	314	29,948
USGS	Church	56,614	314	29,948
USGS	Hospital	3556	314	29,948
USGS	Building	9761	314	29,948
White Pages	Restaurant	825	5	203

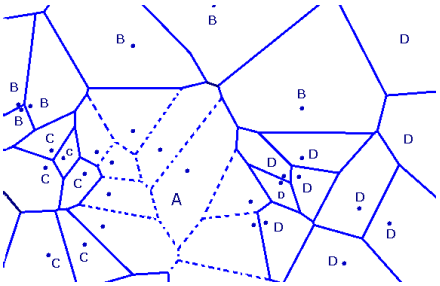

Figure 4. Merging Voronoi cells corresponding to the points with a common feature.

Finally, the areas of each approximate region, the original region and their intersection are computed to measure the precision and recall values. We used approximate region as the *retrieved* set and the original region as the *relevant* set to define our *area-based* precision and recall measures as follows:

$$Precision = \frac{|Retrieved \bigcap Relevant|}{|Retrieved|}$$

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

$$Recall = \frac{|Retrieved \bigcap Relevant|}{|Relevant|}$$

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

where $|A|$ is the cardinality of the set A.

These measures are easily computed using Informix `ST_AREA` and `ST_INTERSECTION` functions that return the area of a polygon and intersection of two polygons, respectively. We refer to the precision-recall measure computed above as the *area-based precision-recall*.

4.2.2 *LDA and QDA*

We used a freely available MATLAB toolbox (Kiefte 2000) for our LDA and QDA implementations. We modified the code to generate the exact boundaries for the approximate map and measure the area-based precision-recall values. In addition, we also measured conventional precision values by classifying a sample of one of the datasets in Table 2 as a test set and performing cross-validation for the training data. However, this measure is not as accurate as our area-based precision-recall measure we defined in Section 4.2.1.

4.2.3 *SVM*

There are several SVM implementations freely available but most of them cannot load our large training datasets. One possible solution to this problem is dividing the original dataset to several chunks of smaller sizes so that each chunk fits into the main memory (Bradley et al. 2002).

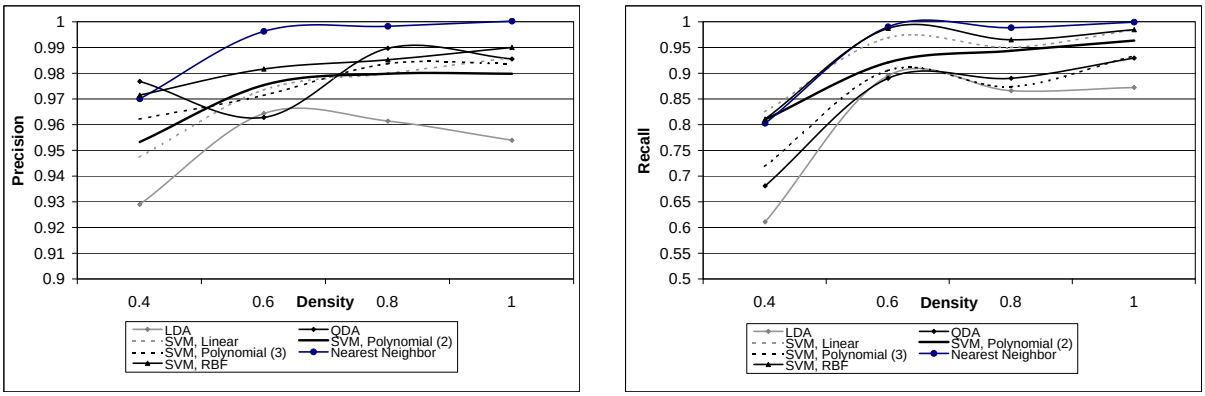

Figure 5: Test-based precision (a) and recall (b) for different methods generating the zip map using USGS data.

RHBNC-SVM (Weston and Watkins 1998) is an open source implementation of SVM that supports multi-class pattern recognition for large datasets using chunking. It enabled us to train several support vector machines for our experiments.

Since we need to generate the best possible trained SVM with the least error, we set the value of the parameter C [1] 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 with large training data, the chunking option was implemented. In our experiments, we trained SVM with four different kernels: radial basis (RBF), linear and polynomial kernels with the degrees of 2 and 3.

4.3 *Results*

In our first set of experiments, we investigated how precisely each classification method can approximate the original map. Figure 5a depicts the precision of four different methods we used to approximate the zip map using the USGS dataset with different densities for the training datasets. We made samples including different subsets of USGS data (see Table 2) as our training datasets with different point densities. Then, we used each method to classify the training datasets and computed the accuracy measures by counting the number of correctly classified data points in our test datasets (i.e. the conventional *test-based* precision measure). As shown in Figure 5a, as the point density in the training data grows, precision of almost all methods increases. Nearest neighbor shows the best accuracy even for low densities. SVM with different kernels generates the second most accurate map. The accuracy of LDA and QDA methods fluctuates over different densities but they create acceptable results with the precision up to 90% for even the sparse training sets. We can also compute the test-based recall in the same way. Figure 5b shows the recall values. Considering the definition of precision and recall, the figure illustrates that although all the methods create good approximations with high precisions but the generated map regions are only small subsets of the original regions when data is sparse. These regions are growing as the density of the training dataset increases.

In the previous experiments, we examined the accuracy of each of the investigated method using uniformly distributed test cases from USGS data. The accuracy of the test-based precision values computed using this approach depends on how well the test datasets can represent the set of all the data points inside each region of the original map. Therefore, we used the features provided by our spatial database system to accurately measure area-based precision-recall values for our classification methods (see Section 4.2.1). Figure 6 depicts area-based precision and recall values for the approximate zip maps generated by our four suggested methods. We used exactly the same map regions generated during the first experiment to compute area-based accuracy measures. As shown in the figure, the higher the number of points in each region, the more accurate the approximate region generated by the method. But the result is quite different from test-based values computed in the previous experiment; area-based values are far less than their correspond-

[1] The bound on the Lagrange multiplier (alpha value) of support vectors.

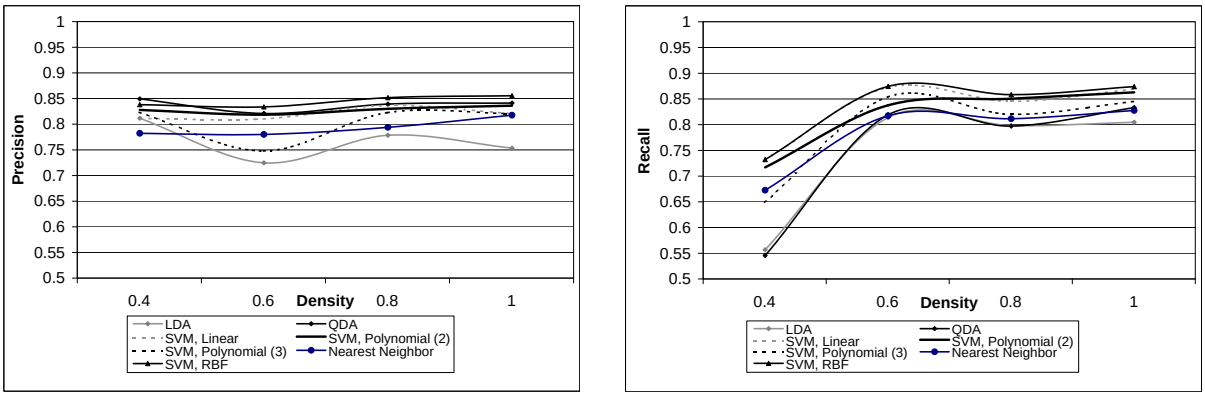

Figure 6: Area-based precision (a) and recall (b) for different methods generating the zip map using USGS data.

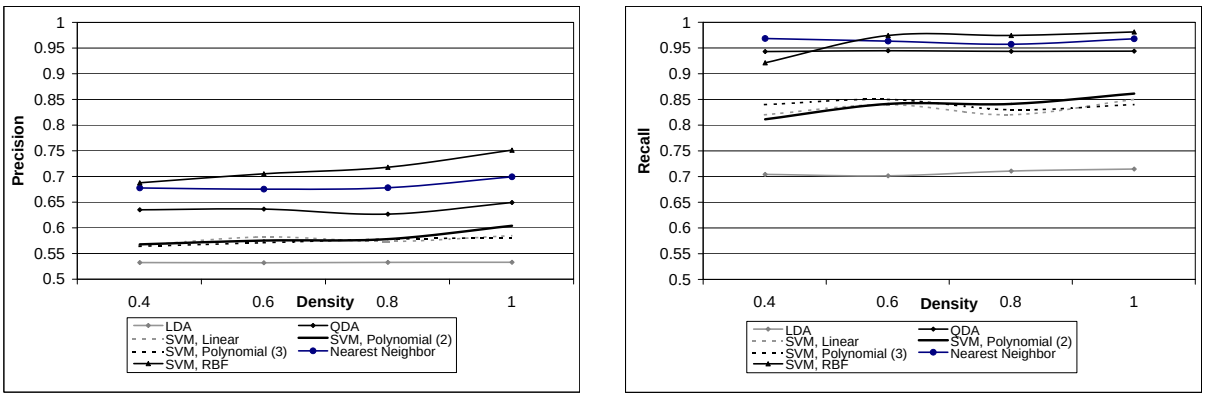

Figure 7: Area-based precision (a) and recall (b) for different methods generating the MSA map using USGS data.

ing test-based values. Even the order in terms of accuracy among different methods has changed. SVM with a radial basis kernel is the superior approach with respect to both precision and recall measures. All other SVM kernels also outperform nearest neighbor and LDA but QDA is comparable to SVM with a polynomial kernel. This set of experiments reveals the fact that the test-based precision-recall measure is not a reliable measure to evaluate the accuracy of different classification methods in generating approximate maps. Instead, the area-based precision-recall measure examines all the false hits and the missing points in a map region and hence is more reliable.

Figure 7 shows area-based precision-recall values for the methods used to generate the approximate MSA map. Considering both precision and recall, SVM with radial basis kernel is still the most accurate method but it results only into a maximum precision of 75% even for dense training data. Nearest neighbor and QDA are in the second place and all other SVM kernels follow them. LDA is the least accurate method in terms of both precision and recall. The intuition here is that the LDA density estimator function defines a density center for each map region which is far from its boundaries in the case of the MSA map with large map regions.

Comparing the two diagrams in Figure 6 and Figure 7 verifies that the zip map generated by each method using a training dataset is more accurate than the MSA map created by the same method using the same training dataset. The reason for this difference is that the regions of the MSA map are much larger than those of the zip map (see Table 1) and the classification method needs different densities to achieve an acceptable approximation for each of these maps.

Our last set of experiments was aimed to study the impact of the training data distribution on the accuracy of the approximate map. We generated two approximate zip maps by training the SVM method using USGS and WP data, respectively. Figure 8 depicts area-based precision-recall values computed for these maps. USGS data is a uniformly distributed dataset but WP data is more dense in the areas close to the center of each zip region. As shown in the figure, considering both precision and recall values, more accurate map regions with small number of training data points

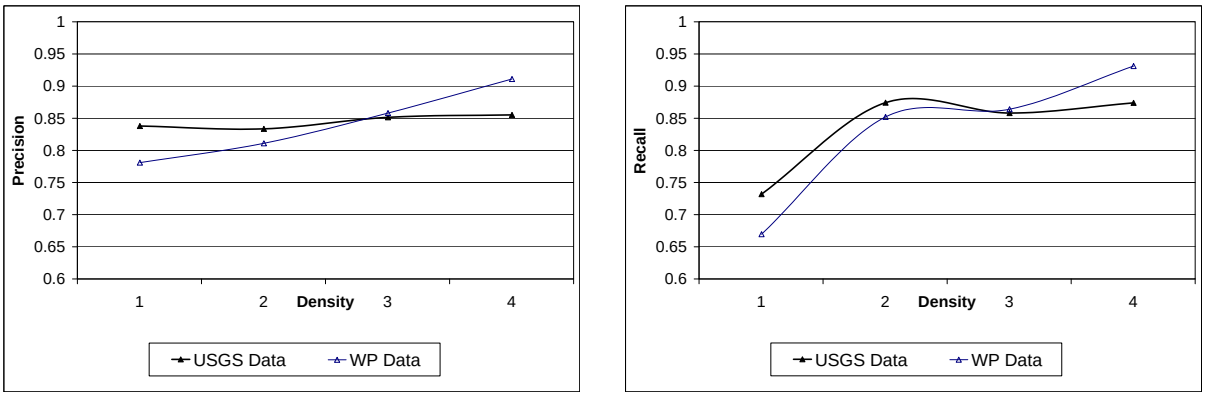

Figure 8: Area-based precision (a) and recall (b) for the zip maps generated by the SVM method with a radial basis kernel using USGS and WP data.

can be generated using USGS data. In contrast, the method trained using WP data outperforms the one trained using USGS data for regions with larger point density. The intuition here is that since the zip map regions are small, using more than 7 training points in those regions that are denser close to their center is sufficient to achieve an acceptable approximate map.

5 RELATED WORK

A relevant body of work in building thematic maps from underlying datasets is the area of spatial interpolation. Spatial interpolation methods use the observation data provided by remote sensing sites or images taken by radars. These methods have been extensively used for generating thematic maps such as land coverage and precipitation maps (Bruin 2000; Dungan 1998; Goovaerts 1999). However, all the studies in this area have focused on different natural phenomena (e.g. vegetation coverage) and tried to find the most accurate map using an environmental dataset. Most of these studies have employed the process models of the phenomena to improve the precision of their interpolation methods. Their approaches are not always applicable to our problem of building thematic maps for general non-natural features (e.g. zip code). The reason is that no process model describing the distribution of these features exists.

Another relevant area consists of regression/classification algorithms that have been proposed in the field of machine learning. These techniques are widely incorporated in numerous research and industrial projects. Comparing to geospatial methods, these methods are model-free. That is, they are general enough to interpolate missing values using only a set of labeled sample data. This feature of the learning algorithms makes them appealing enough to be employed in geospatial-related problems. Hence, we based our study on these machine learning methods and we have already discussed them in details in Section 3.

For the remainder of this section, we briefly describe the main spatial interpolation methods used for mapping natural phenomena. Spatial interpolation is the primary means to estimate values for unmonitored locations. Visualizing the estimated values combined with the set of labeled locations forms the thematic map pertaining to the corresponding feature domain."*Spatial interpolation is the procedure of estimating the values of properties at unsampled sites within an area covered by existing observations*" (Lam 1983). Different spatial interpolation methods have been proposed for environmental datasets with discrete observations at some locations in the environment. These methods are categorized into global and local groups based on the set of observations they use to interpolate missing values. The group of global methods apply a single function to the entire set of observations in the space. *Kriging* is an example of a method in this group. The local methods instead apply a common function repeatedly to subsets of the observed points. These methods such as *Spatial Moving Average (SMA)* and *Triangulated Irregular Network* usually generate the interpolated data as a set of local results.

5.1 *Kriging*

Kriging (Oliver and Webster 1990) is a complicated interpolation technique developed in the field of geostatistics. The technique observes the underlying process in the space using representative

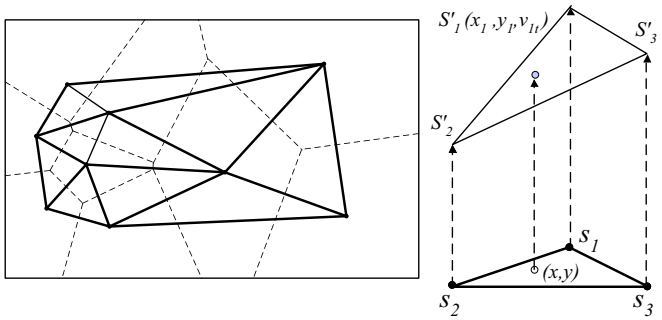

Figure 9. The Delaunay triangulation of the space, interpolating the value of the location (x, y).

variables (e.g. temperature) and computes unknown values of the variable using the values sampled in a limited set of locations. The interpolation method in Kriging is an optimization procedure which uses a model of the process to determine unknown values. This model is given as a variogram of the process. The method assigns optimal weights to the known values in order to predict the unknown values. Kriging is the most extensively used geostatistical interpolation method for predicting values at unrecorded locations. In (Goovaerts 1999), Goovaerts uses three variants of Kriging to incorporate a digital elevation model into the prediction of rainfall. The study reports on the performance of these methods compared to three univariate techniques and concludes that employing a model improve the precision of the interpolation. In (Dungan 1998) Kriging is used to predict the vegetation quantities for an area near the coast of Oregon using samples from a radar image. We cannot use Kriging for building thematic maps of non-natural features such as zip code as there is absolutely no model describing the distribution of these features.

5.2 *Spatial Moving Average (SMA)*

Spatial Moving Average method is widely used in different fields such as GIS and image processing. SMA divides the space using equal size grid cells. The value assigned to each location in the grid cell is then defined as a weighted average of the values of all observation points inside the cell. The corresponding weight of each value is $1/d$ where d is its distance from the center of the grid cell. The method is called *Inverse Square Distance* when the weight of each value is $1/d^2$. In (Goovaerts 1999), Goovaerts reports on interpolating rainfall values using this method as one of the univariate techniques. Our study cannot use SMA because the range of values assigned by SMA is not identical to the discrete finite set of feature values of the labeled observation locations.

5.3 *Triangulated Irregular Network (TIN)*

TIN (Peucker et al. 1978) is a vector-based method used as a digital elevation model. It is a method to generate a 3-dimensional model for the elevation data collected at a set of observation points in 2-d space. The method generates the model in two steps. First, all the observation points which are of the form (x, y, z) are projected to the xy plane. The Delaunay triangulation of the xy plane is created using the set of projected points. This is a unique partitioning of the space using triangles formed by neighboring points in the Voronoi diagram as their vertices. Then, for each triangle in the xy plane ($\triangle s_1 s_2 s_3$ in Fig. 9) the three observation points corresponding to its vertices are considered. Assuming that the points are not collinear, they define a unique 3-d plane. The projection of the triangle to this plane forms a 3-d triangle ($\triangle s'_1 s'_2 s'_3$). Finally, the set of all 3-d triangles defined by the triangles in the Delaunay triangulation is a 3-d visualization of the observation data.

Although TIN is a visualization technique but it has also been used as a spatial interpolation method. Let the z value of each point be the value of the function $f(x, y)$ to be interpolated. To interpolate the value of a location (x, y), first it is located in the set of Delaunay triangles. Then, it is projected to the corresponding 3-d triangle of its surrounding Delaunay triangle. The z value of the projected point is the interpolated value of the location (x, y). Interpolation with TIN assigns values only to the locations inside the *convex hull* of the observation points. That is, it assigns

no value to the locations which are outside of all triangles. This shortcoming of the method is overcome by inserting virtual points on the boundaries of the space. We did not use TIN in our study as the range of values which TIN assigns to the unknown locations is a continuous set and need to be discretisized to be used to build maps of discrete regions.

6 CONCLUSION AND FUTURE WORK

We proposed the use of classification methods to build approximate thematic maps. Through several empirical experiments we identified the accuracy of different methods using the traditional test-based precision measure. We introduced the area-based precision-recall measure, a more accurate measure, and performed different sets 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. The major observations can be summarized as follows:

- Classification methods that generate decision boundaries for all classes can be applied to sample data points to build approximate thematic maps.
- The area-based precision-recall measure verifies that SVM with a radial basis kernel outperforms all the other investigated methods in accuracy.
- The area-based precision-recall values are usually smaller than their corresponding test-based values. Moreover, the area-based measures are more acceptable in practice.
- A spatial database system can be efficiently used to compute the area-based accuracy measures.
- Uniformly distributed features in the training dataset lead to a more accurate map for sparse datastes.

We intend to extract decision boundaries for other classification methods and define new accuracy measures which consider the geometry of the generated map. We also plan to explore more classification techniques and study the way other factors such as the requested resolution for the approximate map impact the accuracy of different classification methods.

ACKNOWLEDGEMENTS

We thank Snehal Thakkar for his help with the experiments on LDA and QDA methods. This research is based upon work supported in part by the National Science Foundation under award numbers IIS-0324955(ITR), EEC-9529152 (IMSC ERC), IIS-0238560 (CAREER), in part by the Air Force Office of Scientific Research under grant numbers F49620-01-1-0053 and FA9550-04-1-0105, by a grant from NASA/JPL, and in part by unrestricted cash gifts from Microsoft. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Barber, C. B., D. P. Dobkin, and H. Huhdanpaa (1996). The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw. 22*(4), 469–483.

Bradley, P., J. Gehrke, R. Ramakrishnan, and R. Srikant (2002). Scaling mining algorithms to large databases. *Commun. ACM 45*(8), 38–43.

Bruin, S. D. (December 2000). Predicting the areal extent of land-cover types using classified imagery and geostatistics. *Remote Sensing of Environment 74*(3), 387–396.

Clarke, K. C. (2002). *Getting Started with GIS* (4th ed.). Prentice Hall.

Dungan, J. L. (1998). Spatial prediction of vegetation quantities using ground and image data. *International Journal of Remote Sensing 19*(2), 267–285.

Goovaerts, P. (1999). Performance comparision of geostatistical algorithms for incorporating elevation into the mapping of precipitation. In *Proceedings of the 4th International Conference on GeoComputation*.

Informix Corporation (2000). Informix spatial datablade module. Version 8.1.

Kiefte, M. (2000). Discriminant analysis toolbox. Version 3.0, ftp://ftp.mathworks.com/pub/contrib/v5/stats/discrim/.

Koperski, K., J. Adhikary, and J. Han (1996). Spatial data mining: Progress and challenges. In *Proceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery*.

Lam, N. (1983). Spatial interpolation methods: A review. *10*(2), 29–149.

Okabe, A., B. Boots, K. Sugihara, and S. N. Chiu (2000). *Spatial Tessellations, Concepts and Applications of Voronoi Diagrams* (2nd ed.). John Wiley and Sons Ltd.

Oliver, M. and R. Webster (1990). Kriging: a Method of Interpolation for Geographical Information Systems. *International Journal Geographic Information Systems 4*(3), 313–332.

Peucker, T. K., R. J. Fowler, and J. J. Little (1978). The triangulated irregular network. In *Proceedings of the ASP-ACSM Symposium on DTM's*.

Tobler, W. (1979). *Cellular Geography, Philosophy in Geography*. Dordrecht: Reidel Publishing Company.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. New York: Springer Verlag.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: John Wiley and Sons.

Verizon Inc. (2004). Verizon SuperPages. http://www.superpages.com/.

Weston, J. and C. Watkins (1998). Multi-class support vector machines. Technical report, Royal Holloway, University of London.