Spatially Adaptive Classification and Active Learning of Multispectral Data with Gaussian Processes

Goo Jun
Dept. of ECE
The Univ. of Texas at Austin
Austin, TX, USA
Email: gjun@mail.utexas.edu

Ranga Raju Vatsavai
GIST/CSE
Oak Ridge National Lab.
Oak Ridge, TN, USA
vatsavairr@ornl.gov

Joydeep Ghosh
Dept. of ECE
The Univ. of Texas at Austin
Austin, TX, USA
Email: ghosh@mail.utexas.edu

***Abstract*—Multispectral remote sensing images are widely used for automated land use and land cover classification tasks. Remotely sensed images usually cover large geographical areas, and spectral characteristics of each class often varies over time and space. We apply a spatially adaptive classification scheme that models spatial variation with Gaussian processes, and apply uncertainty sampling based active learning algorithm to achieve better classification accuracies with a fewer number of samples. The spatially adaptive classifier shows better performances than the conventional maximum likelihood classifier in both passive and active learning settings, and the active learners achieves better classification accuracies than passive learners with fewer number of samples for both classification algorithms.**

***Keywords*-remote sensing; classification; Gaussian process; spatial statistics; active learning**

I. Introduction

A. Land use and land cover classification

In the last couple of decades land use and land cover (LULC) identification with remotely sensed images has become of great interest to researchers from various disciplines including earth scientists and data miners, and it has been applied to a variety of applications such as urban planning, natural resource management, water source monitoring, environmental and agricultural analyses. Remotely sensed multispectral imaging is one of the most widely used technologies for LULC mapping and monitoring, and it provides synoptic and timely information over large geographical areas. Multispectral and hyperspectral image analysis for prediction of LULC classes, however, involves many challenging problems. We investigate solutions for two major problems in this paper: modeling spatial variations within image extents and second minimizing the labeling cost.

B. Modeling spatial variation

In multispectral images, each pixel is represented as a d-dimensional vector, where each element in the vector represents the spectral value taken from the corresponding multispectral band.

With conventional classification methods, it is assumed that spectral signature remains constant across the image. Though this assumption may hold in small spatial footprints, in general spatial signature may not remain constant across the image. Spatial variations in the spectral signature occur due to several reasons: soil type, terrain and climatic conditions. In the presence of spatial variation, a classifier trained on a small region does not generalize well to other areas. Moreover, pooling samples taken across the image may lead to wrong estimation of model parameters. To resolve this problem, a classifier should be able to model underlying variations of spectral information.

Statistical modeling of spatial variation has been well known as spatial statistics or geostatistics. In spatial statistics, nearby points are considered to be more closely correlated to each other than distant points, and this relationship is concisely stated by Waldo Tobler as "Everything is related to everything else, but near things are more related than distant things [1]," also known as the first law of geography. Spatial interpolation techniques for geostatistical data are called kriging [2], where data points are modeled as realizations of underlying spatial random processes. In machine learning, a similar technique has been studied as a Gaussian process regression model, where data points in feature space are modeled as realizations of underlying Gaussian random processes [3]. In this paper, we apply a Gaussian process maximum likelihood (GP-ML) classifier [4] to model spatial variation of remote sensing data, and extend the framework to active learning problems.

C. Minimizing labeling cost

For most supervised machine learning problems, building a good model requires a sufficient number of labeled examples. Often one need 10-30 times the number of dimensions to estimate parameters of statistical distributions [5]. Difficult learning problems with complex decision boundaries require more examples than simpler ones. Often obtaining good ground truth is the most time-consuming and expensive task of the entire learning process, usually involving human experts and additional data. Compared to labeled examples,

unlabeled examples are easier and cheaper to obtain in many cases. Acquiring ground truth for LULC classification is also expensive and time consuming. Remote sensing images taken from satellites easily cover an entire country, or a continent. Obtaining highly reliable class labels for all regions covered by these images is practically not possible. In contrast, we can easily obtain billions of unlabeled data from these images. Active learning algorithms are designed to minimize the number of labeled samples to achieve desired level of accuracies. We will exploit active learning to further enhance the spatially adaptive classification of remote sensing imagery.

II. Related work and our contribution

Spatial variation of remote sensing data has been studied by many researchers. Atkinson and Lewis provided a survey of geostatistical methods for remote sensing classification [6], but most of these methods are about spatial smoothing and weighting. In [7], authors applied geostatistical analysis of hyperspectral data but did not provide tools for classification. Goovaerts combined spectral classifier with spatially modeled prior probabilities using indicator kriging [8].

Active learning has been a popular topic in the machine learning community for last two decades. One of the most well-known active learning algorithm is Query-by-committee (QBC) [9], where a committee of independent classifiers chooses samples to be queried. MacKay [10] proposed an active learning framework where the learner chooses an example that has the most expected information gain. Cohn *et. al.* [11] proposed a method based on a statistical analysis to select the sample that minimizes the variance of a given model. Lewis and Gale [12] proposed a sampling criterion for active learning called uncertainty sampling, which we employ in this paper. Since uncertainty sampling does not refer to a single uncertainty measure nor a single classification algorithm, various kinds of uncertainty sampling strategies can be used depending on problem domains. An active sampling strategy for Gaussian process regression is studied by Krause and Guestrin [13], but their work is on reducing uncertainties of the regression model itself and is not intended to minimize classification errors.

There are active learning algorithms developed for remote sensing problems: Rajan *et al* [14] provided a framework for active learning under spatial and temporal variations of hyperspectral data, and Jun and Ghosh [15] extended this framework to incorporate transfer learning techniques. Liu *et al* [16] proposed an active learning algorithm that models the label acquisition cost as a sum of distances travelled to visit the acquired points. In these works, however, classifiers do not utilize spatial relations between samples at different locations.

In this study we present a spatially adaptive classification method based on the GP-ML framework [4] and apply it to the classification of multispectral data that has relatively large spatial extents compared to other remotely sensed images. We further extend this classification scheme by integrating active learning principles to rank the unlabeled samples using uncertainty measures. We have conducted several experiments which show the benefits of our algorithm over the conventional maximum likelihood classifier.

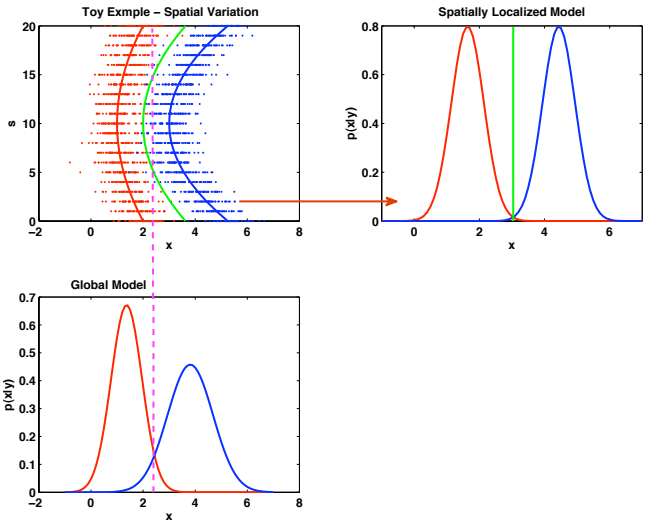

Figure 1. Toy example of spatially varying distribution. Upper left: distribution of randomly generated points according to Gaussian distributions with spatially varying means. Upper right: Gaussian distributions at a single location. Lower left: globally estimated Gaussian distributions without spatial information.

III. GP-ML

A. Statistical Framework

The conventional maximum-likelihood classifier (MLC) typically models the class-conditional distribution, $p(\mathbf{x}|y)$, as a multi-variate Gaussian distribution:

$$p(\mathbf{x}|y=i) \sim N(\boldsymbol{\mu_i}, \Sigma_i) \ , \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_d)^T$ is a d-dimensional vector representing spectral bands of a pixel, and $y \in \{1, 2, ..., c\}$ is the LULC class label. Parameters for multi-variate Gaussians, $\boldsymbol{\theta_i} = (\boldsymbol{\mu_i}, \Sigma_i)$, are obtained by a maximum-likelihood estimation, and assumed to be constant over all possible locations. As discussed earlier, this assumption does not hold in general.

Figure 1 demonstrates a toy example of spatially varying distributions. Red and blue points in the upper left plot indicate two different classes. The x-axis indicates feature values (one-dimensional feature), and the y-axis indicates spatial coordinates (one-dimensional space). At a given location, data points of each class are randomly generated from a univariate Gaussian distribution. Means of class conditional distributions are generated by a smooth quadratic

Algorithm 1 Outline of the GP-ML algorithm

Input: Traning data $\{(\mathbf{x}_1, \mathbf{s}_1, y_1), (\mathbf{x}_2, \mathbf{s}_2, y_2), ..., (\mathbf{x}_n, \mathbf{s}_n, y_n)\}$, test instance $(\mathbf{x}^*, \mathbf{s}^*)$. $\mathbf{x} \in R^d,\ \mathbf{s} \in R^2,\ y \in \{1, 2, ..., c\}$.

1: Estimate hyperparameters $(\lambda,\ \sigma_{ij}^2),\ 1 \le i \le c,\ 1 \le j \le d$.
2: **for** $i = 1$ to c **do**
3: $(X_i, S_i) \leftarrow \{(\mathbf{x}_k, \mathbf{s}_k) | y_k = i\}$.
4: $\mathbf{x}_i^j \leftarrow$ column vector with j-th bands of $\forall \mathbf{x} \in X_i$.
5: Estimate predictive mean at $\mathbf{s}^*$, $\boldsymbol{\mu}_i(\mathbf{s}^*) = (\mu_{i1}(\mathbf{s}^*), \mu_{i2}(\mathbf{s}^*), ..., \mu_{id}(\mathbf{s}^*))$:

$$\mu_{ij}^*(\mathbf{s}^*) = K_j^i(\mathbf{s}^*, S)[K_j^i(S, S)]^{-1}\mathbf{x}_i^j \ .$$

6: Estimate covariance matrix Σ_i from X_i.
7: Calculate $P_i = P(y = i)p(\mathbf{x}^*(\mathbf{s}^*)|y = i)$, where $p(\mathbf{x}^*(\mathbf{s}^*)|y = i) \sim N(\boldsymbol{\mu}_i^*(\mathbf{s}^*), \Sigma_i)$.
8: **end for**

Output: $y^* = \arg\max_i P_i$.

function of spatial coordinates, and 21 different Gaussian distributions are plotted for each class, for $s = 0, 1, ...20$. Standard deviations are assumed to be constant, and to be identical for both classes. The red and blue curves indicate actual means used to generate red and blue data points, respectively. The upper right figure shows class conditional distributions of two classes for $s = 3$, while the lower left plot shows Gaussian distributions obtained by ML estimation without any spatial information. The green curve indicates the ideal (Bayesian) decision boundary, and the dashed line in magenta indicates the decision boundary obtained by a global MLC. As can be seen in the example, it is possible to achieve better classification accuracies by proper modeling of spatially varying parameters.

In the previous example, the underlying quadratic model of spatial variation was already known. In most real situations, we do not have such knowledge. Since the spectral characteristics of a class in a multispectral image are influenced by several factors, assuming a single parametric model is not desirable. Instead, we employ a non-parametric Gaussian process model. In the GP-ML framework, the class-conditional distribution of the i-th class is modeled as a function of spatial coordinate $\mathbf{s}$:

$$p(\mathbf{x}(\mathbf{s})|y = i) \sim N(\boldsymbol{\mu}^i(\mathbf{s}),\ \Sigma^i) \ . \tag{2}$$

where $\boldsymbol{\mu^i(s^*)} = (\mu_1^i(\mathbf{s}^*), \mu_2^i(\mathbf{s}^*), ..., \mu_d^i(\mathbf{s}^*))$. Each spectral band of data from from the i-th class is modeled as a random process indexed by a spatial coordinate $\mathbf{s} = (s_1, s_2)$; hence the j-th band of $\mathbf{x}^i$, x_j^i, can be written as

$$x_j^i(\mathbf{s}) = f_j^i(\mathbf{s}) + \epsilon_j^i \ , \tag{3}$$

where $f_j^i(\mathbf{s})$ is a Gaussian random process and ϵ_j^i is an additive white Gaussian noise (AWGN):

$$\epsilon_j^i \sim \mathcal{N}(0, \sigma_{ij}^2) \ .$$

Given $f_j(\mathbf{s})$, then the class conditional distribution of x_j is

$$p(x_j^i(\mathbf{s})|f_j^i(\mathbf{s})) = \mathcal{N}(f_j^i(\mathbf{s}),\ \sigma_{ij}^2) \ .$$

We assume a (zero-mean) Gaussian process for $f_j(\mathbf{s})$:

$$f_j^i(\mathbf{s}) \sim \mathcal{GP}(0,\ K_j^i(\mathbf{s}_l, \mathbf{s}_m)) \ ,$$

where $K_j(\mathbf{s}_l, \mathbf{s}_m)$ is a spatial covariance function between locations $\mathbf{s}_l$ and $\mathbf{s}_m$. The zero-mean prior assumption corresponds to the simple kriging model in spatial statistics [2]. In practice, we can approximately satisfy the zero-mean assumption by normalization.

In the GP-ML model, we only model spatial variation of the mean parameters, $\boldsymbol{\mu_j}$, and the spectral covariance Σ is assumed to be constant without any spatial variation. Each band is assumed to be independent of each other in spatial sense, which means we do not consider cross-correlation of $x_j(\mathbf{s}_l)$ and $x_k(\mathbf{s}_m)$ for $j \neq k$ and $\mathbf{s}_l \neq \mathbf{s}_m$. This assumption is generally not true, but modeling spatial cross-correlation makes the model too complex. In spatial statistics, the technique of modeling multiple correlated target variables is called *cokriging* [2]. In cokriging, every target variable is modeled as a linear combination of all other target variables at all other locations, and we need $d \times d$ covariance functions to model these linear relations. The cokriging model is usually impractical and too demanding[8], since it requires solving $(n + 1) \cdot d$ linear equations for n data points with d dimensions. In addition to the computational burden, it makes the system too sensitive to the noise leading to inaccurate parameter estimates.

Assuming zero-mean Gaussian priors for $f_j^i(\mathbf{s})$, predictive distribution for a new location $\mathbf{s}^*$ is easily obtained from the formula for conditional distributions of jointly Gaussian random vectors [3] as described in Algorithm 1. Characteristics of a Gaussian random process is solely defined by a covariance function. We used the isometric squared exponential covariance function [3] in our model. Using the squared exponential covariance function, the covariance of the j-th bands at locations $\mathbf{s}_l$ and $\mathbf{s}_m$ is modeled as:

$$K_j^i(\mathbf{s}_l, \mathbf{s}_m) = \exp\left(-\frac{||\mathbf{s}_l - \mathbf{s}_m||^2}{2\lambda^2}\right) + \sigma_{ij}^2\delta_{lm} \ , \tag{4}$$

where δ_{lm} is the Kronecker delta function.

B. Hyperparameters

1) Length Parameter: Although a Gaussian process model is often told to be nonparametric, the nature of a

Gaussian process is governed by several hyperparameters. The squared exponential covariance function in (4) has two hyperparameters: the noise power σ_{ij}^2, and thelength parameter λ. We assume unit-variance random processes, assuming pre-normalized data. A length paramter determines how fast the correlation between two points decreases as the distance between the points increases. We used 10-fold cross-validation to find the best length parameter. Unlike [4], we randomly selected training and validation sets since our data points are already scattered over the image. A wide range of length parameters is tested and the parameter that yields best overall classification accuracy is selected.

2) Noise power: The noise parameter σ_{ij}^2 determines how tightly should the Gaussian process be fitted to the given points. In kriging literature, effects from the noise are referred as *nugget effects* [2], and the name came from the mining terminology for which kriging was originally developed. Unlike kriging, where embedding actual noise model using nugget effect is very indirect and hard to analyze, the Gaussian process formulation enables including the exact AWGN term and provides Wiener-filtered processes. In our experiments, we estimated the noise power using localized means. Our data points are sampled in 3×3 grids for each spatial locations, hence we approximately assumed the fitted Gaussian process to be constant in the grid. This approximation holds because the size of the grid is negligibly small compared to the distances between sampled locations and the length scale parameter. For each grid, localized means are obtained by averaging feature values in the grid, and then σ_{ij}^2 is calculated as averaged squared differences of spectral values from the localized means.

IV. Active learning

In machine learning literatures, active learning refers to learning algorithms where the learner actively chooses its own training set, and it is different from passive learning algorithms where the learner is trained with given (often randomly selected) dataset and has no control on how the training set is constructed. Typically active learning algorithms consist of several steps. Initially a learner is trained on a small labeled dataset, and then the learner is exposed to a pool or stream of unlabeled data. The learner chooses k examples those are considered most useful from the unlabeled data, and acquires ground truth for them. Then the learner is re-trained using additional labeled data, and the choose-and-learn process is repeated. In a slightly different setting of query-based active learning, a learner could generate unlabeled examples on its own instead of selecting from given set of data. In most cases, the goal of an active learning algorithm is either achieving a lower error rate than passive learning algorithms with the same number of labeled samples, or achieving equal error rate with a fewer number of labeled samples, if not both.

Algorithm 2 Uncertainty sampling

Input :
Labeled set D_L
Unlabeled set D_{UL}

1: **while** $|D_{UL}| > 0$ **do**
2: Train the classifier with D_L to get

$$P_i(\mathbf{x}(\mathbf{s})) = P(y = i)p(\mathbf{x}(\mathbf{s})|y = i),\ \forall(\mathbf{x}, \mathbf{s}) \in D_{UL}\ .$$

3: Get posteriors $\forall i, \forall(\mathbf{x}, \mathbf{s}) \in D_{UL}$:

$$P(y = i|\mathbf{x}(\mathbf{s})) = \frac{P_i(\mathbf{x}(\mathbf{s}))}{\sum_{l=1}^{c} P_l(\mathbf{x}(\mathbf{s}))}\ .$$

4: Get uncertatinty scores $\forall(\mathbf{x}, \mathbf{s}) \in D_{UL}$:

$$u(\mathbf{x}(\mathbf{s})) = \frac{1}{P(y = i|\mathbf{x}(\mathbf{s})) - P(y = j|\mathbf{x}(\mathbf{s}))}\ ,$$
$$\text{where}\quad i = \arg\max_i P(y = i|\mathbf{x}(\mathbf{s}))\ ,$$
$$j = \arg\max_{j \neq i} P(y = j|\mathbf{x}(\mathbf{s}))\ .$$

5: Pick $(\mathbf{x}_p, \mathbf{s}_p) = \arg\max u(\mathbf{x}(\mathbf{s})), (\mathbf{x}_p, \mathbf{s}_p) \in D_{UL}$.
6: Query class label y_p for $(\mathbf{x}_p, \mathbf{s}_p)$.
7: Update labeled and unlabeled sets:

$$D_L \leftarrow D_L \cup \{(\mathbf{x}_p, \mathbf{s}_p, y_p)\}$$
$$D_{UL} \leftarrow D_{UL}/\{(\mathbf{x}_p, \mathbf{s}_p)\}$$

8: **end while**

Active learning is a useful technique when we have ample amounts of unlabeled data, but class labels are expensive to obtain. Different active learning algorithms use different criteria for judging usefulness of unlabeled points, and it results in different selections. One of the most popular approaches for active learning is the loss-reduction method, where each unlabeled point is evaluated by the expected decrease of the loss function. The loss-reduction type approach generally requires re-training of the learner with all members (or randomly selected subset) of the unlabeled set; hence it is computationally expensive.

Another popular approach is uncertainty sampling [12], where the learner chooses samples that it is most uncertain about their class labels. Uncertainty sampling based algorithms do not require re-training of the learner for every unlabeled example to be evaluated, and can be easily incorporated with many classification algorithms. For example, the well-known Query by Committee (QBC) [9] algorithm also can be thought as a kind of uncertainty sampling strategy, where disagreement between committee members is used as an uncertainty measure.

Methods described in [14] use loss-reduction-based active learning for classification of hyperspectral data, where the loss function is the expected KL-divergence of the unlabeled point. An uncertainty sampling based active learning for hyperspectral data as well is presented in [16]. In both active learning algorithms, estimated posterior probability is used as a criterion to assess unlabeled data, and we need an

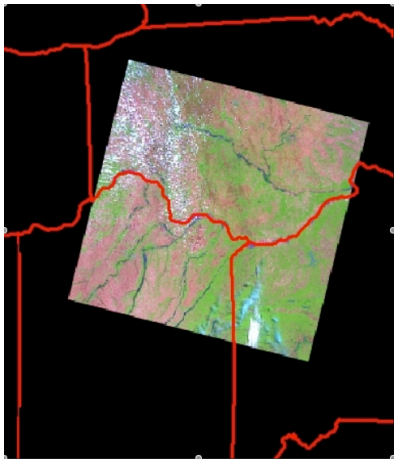

Figure 2. AWiFS False Color Composite Image

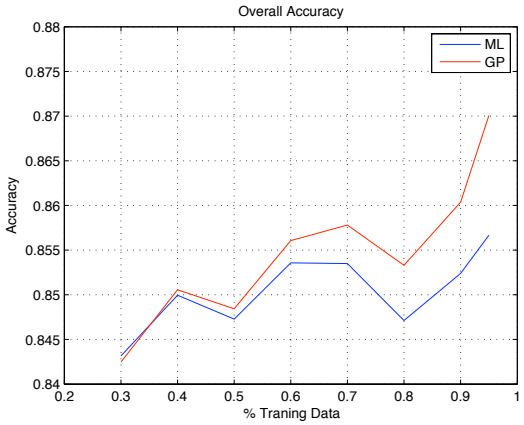

Figure 3. Learning curves of ML and GP-ML classifiers

accurate initial model that can evaluate usefulness of the unlabeled points for efficient active learning. In the presence of spatial variation, using spatially adaptive classification methods such as GP-ML algorithm provides the learner better estimates for usefulness of unlabeled examples. The uncertainty measure we used in this paper is the same as in [16], and it is defined as the inverse of the differences between posterior probabilities of the most probable and the next probable class labels. Outline of the active learning algorithm used in this paper is described in Algorithm 2.

V. Experiments

A. Data

We used an early summer Advanced Wide Field Sensor (AWiFS) data acquired by Resourcesat-1 satellite on June 16, 2008. The image has 8495 pixels per line and there are 8488 lines. Pixel ground resolution is 56 x 56 meters and the total image width is 370 Km. On the other hand, width of most widely used Landsat TM image is about 180 Km and pixel width is 30 meters. Given large spatial extents of AWiFS data, it is clear that spatial variation is much higher than the Landsat images. As shown in Figure 2, the image covers eastern part of Iowa and western part of Illinois. Corn and Soybean are dominant classes. Other classes considered are Pasture/Hay, Water, Deciduous Forest, and Urban (developed areas). We have collected 3,573 labeled samples from 397 spatial locations (3×3 samples per each location). Six different LULC classes are used as class labels. We now describe each experiment in detail.

B. GP-ML model

To evaluate performances of the GP-ML classifier, we varied the portion of randomly selected training data. Results are shown in Figure 3. The x-axis shows fractions of the training data with respect to the entire data, and the remaining portions are used as test sets. Each experiment is repeated 100 times. Since some classes have very small number of samples, stratified sampling is done to prevent empty classes. When training data has only 30% of the entire data, the GP-ML classifier does not show better results than the ML classifier, but as the size of the training set increases, we can observe that the gap between the GP-ML and ML results is widening. The result implies that for proper modeling of spatial variation, we need a good amount of labeled data, addressing the role of active learning to maximize performances with a fewer number of labeled samples.

Figure 5 shows how we predicted spatially varying means with the Gaussian process model. Plots are generated by taking a narrow strip of the given image and including all points in the strip to obtain a one-dimensional Gaussian process regression. Blue curves are the means modeled by Gaussian processes, and straight green lines are the global mean estimated by the ML classifier. Blue dots indicate training data (averaged for each 3×3 grid), and red ones are test data (all 9 points are plotted). The figure is generated by using 70% of the entire data as training data.

C. Active learning

Active learning results are shown in Figure 4. For active learning experiments, we used 10-fold cross validations to construct common test sets for all four different classifiers, and then subsampled the training set into initially labeled data and initially unlabeled data. Active and random selection of data points are done within the pre-defined unlabeled data, without altering the test set. Picking points from the test set usually produces unrealistically high accuracies for uncertainty sampling algorithms, because the learner always picks the most uncertain points. As a result, the remaining test set eventually tends to have only easy test cases. Initially all classifiers are trained using 30% of the each training set of the cross-validation setup, and then picks

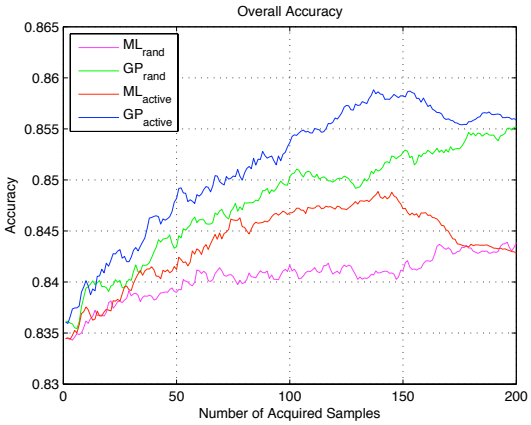

Figure 4. Active learning results

one point at each round until all points in the unlabeled dataset are picked. The number 30% is determined from the experimental results above, because that is the point where the GP-ML and ML classifiers have similar performances. We repeated 10-fold cross validation 10 times, and the results from 100 experiments are averaged. Results in Figure 4 shows the typical banana curves for passive and active learners. Because both active and passive learners initially trained with the same training set, the curves start at the same point. Two curves always meet at the end when we have a finite number of unlabeled examples, because adding all points from the unlabeled dataset again construct the same training set again. In real experiments where the number of unlabeled data is virtually infinite, the curves are not guaranteed to meet each other after the starting point. As shown in the plots, GP-ML classifiers dominate ML classifiers at all stages. Comparing active learners to passive learners, it is easily recognizable that the active GP-ML classifier achieves better accuracies than the passive one, and the same statement holds for the active and passive ML classifiers.

VI. Conclusion and Future Works

In this paper, we compared classification accuracies of the GP-ML classifier to those of the ML classifier. The spatially adaptive GP-ML classifier shows better performances than the conventional maximum likelihood classifier in both passive and active learning settings, and the active learners achieve better classification accuracies than passive learners with fewer number of samples for both classification algorithms. It is also observable that the gain from spatial model become greater when we have more training samples. Although we applied the GP-ML framework only to spatially varying data, the algorithm can be easily extended to spatio-temporal datasets. It is also notable that proposed algorithms can be easily combined with loss-reduction based active learning algorithms, too. In our follow-up research, we will investigate these possible extensions of the proposed framework.

VII. Acknowledgments

We would like to thank Dr. Budhendra Bhaduri, Eddie Bright, Anil Cheriyadat, and Chris Symons of ORNL for giving inputs into this research. Prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UT-Battelle, LLC for the U. S. Department of Energy under contract no. DEAC05-00OR22725. This research is supported by LDRD grant from ORNL. This research is also partly supported by NSF Grant IIS-0705815. We would like to thank our collaborators G. Bethel, R. Tetrault of USDA, S. Johnson of Global Marketing Insights, and B. Doorn of NASA for guidance and research inputs which have greatly helped us in shaping this research. AWiFS data for this research was kindly provided by the USDA.

References

[1] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic Geography*, vol. 46, pp. 234–240, 1970. [Online]. Available: http://www.jstor.org/stable/143141

[2] N. Cressie, *Statistics for Spatial Data*. Wiley, New York, 1993.

[3] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[4] G. Jun and J. Ghosh, "Spatially adaptive classification of hyperspectral data with gaussian processes," in *IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS 09)*, 2009.

[5] P. M. Mather, *Computer processing of remotely-sensed images: an introduction*. New York, NY, USA: John Wiley & Sons, Inc., 2004.

[6] P. M. Atkinson and P. Lewis, "Geostatistical classification for remote sensing: an introduction," *Comput. Geosci.*, vol. 26, no. 4, pp. 361–371, 2000.

[7] D. A. Griffith, "Modeling spatial dependence in high spatial resolution hyperspectral data sets," *Journal of Geographical Systems*, vol. 4, no. 1, pp. 43–51, 2002. [Online]. Available: http://dx.doi.org/10.1007/s101090100073

[8] P. Goovaerts, "Geostatistical incorporation of spatial coordinates into supervised classification of hyperspectral data," *Journal of Geographical Systems*, vol. 4, no. 1, pp. 99–111, 2002. [Online]. Available: http://dx.doi.org/10.1007/s101090100077

[9] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh, PA, USA: ACM Press, 1992, pp. 287–294.

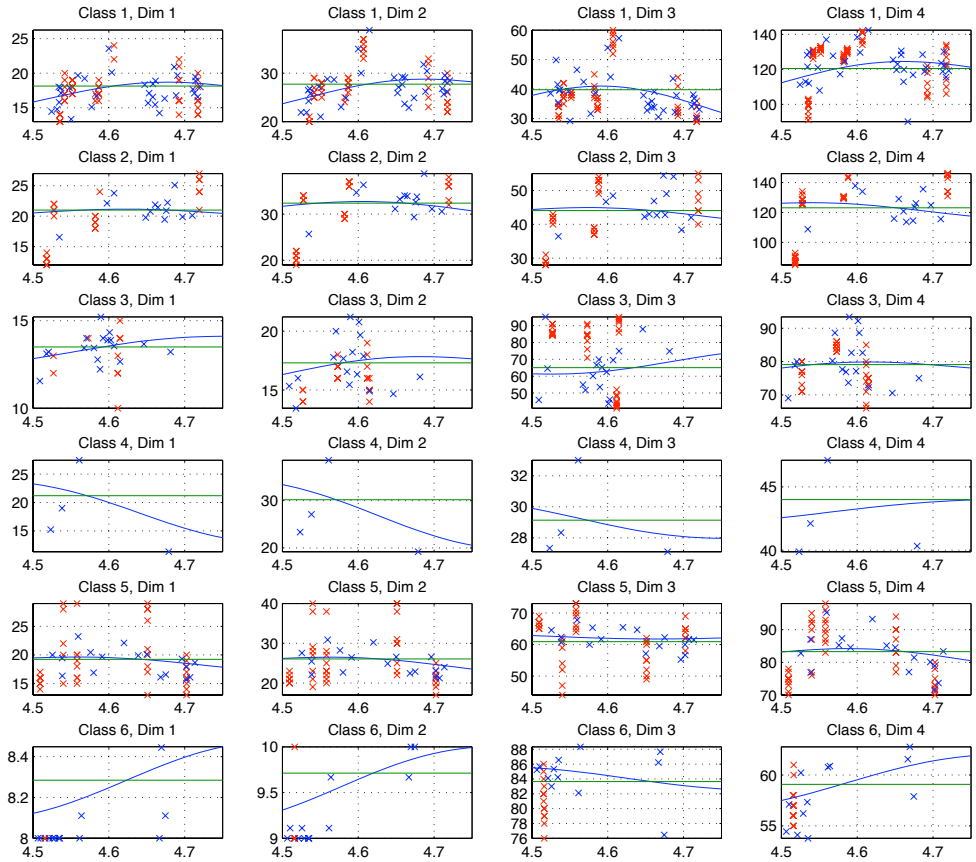

Figure 5. Means modeled by Gaussian process along one direction. Blue points are training data, and red points are test data. The straight green line indicates constant mean by ML estimation.

[10] D. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992. [Online]. Available: citeseer.ist.psu.edu/47461.html

[11] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artifical Intelligence Research*, vol. 4, p. 129, 1996. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:cs/9603104

[12] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Springer-Verlag New York, Inc., 1994, pp. 3–12.

[13] A. Krause and C. Guestrin, "Nonmyopic active learning of gaussian processes: an exploration-exploitation approach," in *ICML '07: Proceedings of the 24th international conference on Machine learning*. New York, NY, USA: ACM, 2007, pp. 449–456.

[14] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, 2008.

[15] G. Jun and J. Ghosh, "An efficient active learning algorithm with knowledge transfer for hyperspectral data analysis," in *IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS 08)*, 2008.

[16] A. Liu, G. Jun, and J. Ghosh, "Spatially cost-sensitive active learning," in *Proceedings of 2009 SIAM International Conference on Data Mining (SDM 09)*, 2009.