SPATIALLY ADAPTIVE CLASSIFICATION OF HYPERSPECTRAL DATA WITH GAUSSIAN PROCESSES

Goo Jun *Joydeep Ghosh*

Department of Electrical and Computer Engineering
The University of Texas at Austin, Austin TX 78712, USA
{gjun, ghosh}@mail.utexas.edu

ABSTRACT

Automated classification of land cover types based on hyperspectral imagery often involves a large geographical area, but class labels are available for only small portions of the entire area. Moreover, the spectral signature of the same land cover class may vary substantially over different locations. When a classifier is trained on a specific geographical location and applied to other areas, it often performs poorly because of such spatial variation of spectral signatures. In this paper, we propose a novel framework for classification of hyperspectral data: a Gaussian-Process Maximum-Likelihood (GP-ML) model where the mean of each spectral band is spatially modeled using a Gaussian process. Our framework provides a practical and effective way to model spatial variations of high dimensional data such as hyperspectral images for classification problems.

***Index Terms*—** hyperspectral data, remote sensing, classification, Gaussian process, kriging, spatial information

1. INTRODUCTION

In recent years, land cover classification by hyperspectral image (HSI) data analysis has become an important part of remote sensing research [1]. Compared to conventional multi-spectral images where each pixel usually contains tens of bands, pixels in hyperspectral image usually consist of more than a hundred spectral bands, providing fine-resolution spectral information. HSI can cover very large areas, but it is not usually possible to obtain highly accurate class labels for all locations in the image. Figure 1 shows the Botswana image used in our experiments with its land cover class map. Different colors indicate different individual land cover classes. The gray areas in the figure denote areas without any class label. As shown in the figure, only a small fraction of the entire region actually has class labels. Because we have spatially restricted training data, we need a classification system that can accurately predict class labels for test data that is spatially distant from the training data. Figure 2 shows how spectral signatures of a single class change over different spatial locations. Predicting land cover classes under spatial variation is a challenging problem.

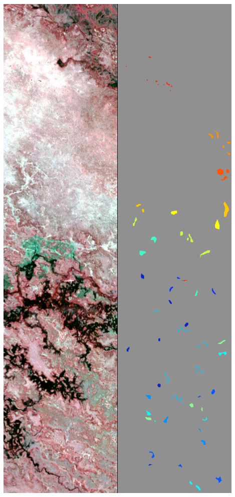

Fig. 1. Botswana image with a class map

Spatial modeling of data has long been studied as an important field of statistics, called spatial statistics or geostatistics. The first law of geography according to Waldo Tobler is “Everything is related to everything else, but near things are more related than distant things [2].” This well describes the importance of neighborhood information as well as global (non-myopic [3]) relationships between data points. Kriging [4] is a well-known technique to model spatial dependencies of data points, and it has been widely studied for various problems of spatial statistics. In kriging, each point is modeled as an outcome of a random process, and this approach has recently been adopted by the machine learning community through its interest in Gaussian process models [5]. In a Gaussian process model, data points are modeled as realizations from a Gaussian random process prior. Prediction for a new sample is obtained by calculating posterior distributions from observed data points and a covariance function. Gaussian process and kriging are closely related. Both techniques

This work was supported by NSF Grant IIS-0705815.

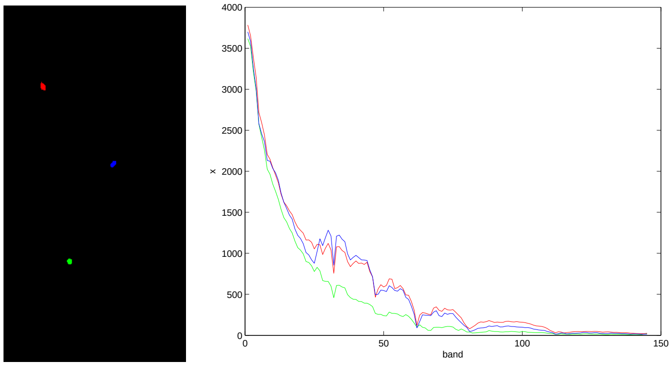

Fig. 2. Spectral signatures of water class at different locations

share important key concepts, but differ in some details. Most of all, Gaussian process models in machine learning usually work in feature spaces, while kriging mostly deals with physical spaces. We exploit both techniques in our experiments by applying Gaussian process models to model spectral variations over physical spaces.

2. RELATED WORKS

There have been algorithms developed for hyperspectral data that are intended for spatially distant datasets, or that use spatial information. For example, Rajan *et al* [6] provided a framework to transfer knowledge between different spatial and temporal locations, but this approach does not utilize spatial relations between locations. Another approach is to add information from homogeneous neighborhoods as in [7], but it does not model varying spectral signatures of hyperspectral data directly.

Griffith analyzed spatial dependencies of hyperspectral data [8], but did not provide classification methods. Atkinson and Lewis surveyed geostatistical methods for remote sensing, and introduced several spatial smoothing and weighting methods [9]. The closest approach to this paper is by Goovaerts[10], where the prior probability of the i-th class, $P_i(\mathbf{s})$, is modeled by indicator kriging. Gaussian process has been long known in spatial statistics as *kriging* [4], but kriging has been considered to be only suitable for modeling of single or small number of target variables. In this paper, we directly model spatially adaptive class-conditional distributions for each band.

3. METHODS

3.1. GP-ML framework

Let $\mathbf{x} = (x_1, x_2, ..., x_d)^T$ be a d-dimensional vector representing spectral bands of a pixel in a hyperspectral image, and $y \in \{y_1, y_2, ..., y_c\}$ be a class label that indicates land cover type. The class-conditional probability distribution $p(\mathbf{x}|y_i)$ is usually assumed to be multivariate Gaussian:

$$p(\mathbf{x}|y_i) \sim N(\boldsymbol{\mu_i}, \Sigma_i) \ , \quad (1)$$

where $\boldsymbol{\mu_i}$ is the mean vector and Σ_i is the covariance matrix of the i-th class. For simple notation, let us focus for now on a single class and omit i. Typically, both $\boldsymbol{\mu}$ and Σ are considered to be constant over the entire image. Instead we model x_j, the j-th band of $\mathbf{x}$, as a random process indexed by a spatial coordinate $\mathbf{s} = (s_1, s_2)$:

$$x_j(\mathbf{s}) = f_j(\mathbf{s}) + \epsilon_j \ ,$$

where $f_j(\mathbf{s})$ is a Gaussian random process and ϵ_j is an additive white Gaussian noise (AWGN) term, $\epsilon \sim \mathcal{N}(0, \sigma^2_{\epsilon_j})$. Our prior for $\mathbf{f}(\mathbf{s})$ is a (zero-mean) Gaussian process:

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

where $K_j(\mathbf{s}_l, \mathbf{s}_m)$ is a spatial covariance function between locations $\mathbf{s}_l$ and $\mathbf{s}_m$. Then given $f_j(\mathbf{s})$, the distribution of x_j is also Gaussian:

$$p(x_j(\mathbf{s})|f_j(\mathbf{s})) = \mathcal{N}(f_j(\mathbf{s}), \ \sigma^2_{\epsilon_j}) \ .$$

We assume each band is spatially independent of each other, hence neglecting 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 implies that the spectral covariance Σ is assumed to be constant without spatial variation. Modeling multiple correlated target variables has been studied in spatial statistics, and it is called *cokriging* [4]. It is impractical and too demanding, however, to model hyperspectral data directly by cokriging [10], since cokriging requires solving $(n + 1) \cdot d$ linear equations for n data points with d dimensions, and it makes the matrix so big that the system becomes sensitive to noise and inaccurate parameters.

Assume now that we have a set of labeled data points from this class, $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$, located at corresponding spatial coordinates $S = (\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n)$, and let $\mathbf{x}_j$ be a vector consists of j-th bands in X, $\mathbf{x}_j = (x_{1j}, x_{2j}, ...x_{nj})^T$. Then the predicted distribution of the j-th band of a new data point $\mathbf{x}^*$ at coordinate $\mathbf{s}^*$ can be easily derived from the conditional distribution of jointly Gaussian random vectors. The distribution of x^*_j given $\mathbf{x}_j$ and S is Gaussian with the mean:

$$\mu_j(\mathbf{s}^*) = K_j(\mathbf{s}^*, S)[\Sigma_S + \sigma^2_{\epsilon_j} I]^{-1}\mathbf{x}_j \ , \quad (2)$$

From (2), we can derive our spatially adaptive class-conditional Gaussian distribution:

$$p(\mathbf{x}(\mathbf{s}^*)|y_i) \sim N(\boldsymbol{\mu_i}(\mathbf{s}^*), \ \Sigma_i) \ , \quad (3)$$

where $\boldsymbol{\mu_i}(\mathbf{s^*}) = (\mu_{i1}(\mathbf{s}^*), \mu_{i2}(\mathbf{s}^*), ..., \mu_{id}(\mathbf{s}^*))$. Maximum-likelihood classification is done with spatially adapted Gaussian distributions to find y_i that maximizes the posterior probability $p(y_i|\mathbf{x}(\mathbf{s}))$.

The popular squared exponential covariance function is employed [5]. The covariance function is assumed to be identical over all classes, and over all bands except for a noise term:

$$K_j(\mathbf{s}_l, \mathbf{s}_m) = \exp\left(-\frac{||\mathbf{s}_l - \mathbf{s}_m||^2}{2\lambda^2}\right) + \sigma^2_{\epsilon_j}\delta_{lm} \ , \quad (4)$$

where δ_{lm} is the Kronecker delta function.

3.2. Fitting Hyperparameters

In the Gaussian process model, a covariance function determines the nature of the process, and the covariance function is characterized by hyperparameters. In (2), we have two hyperparameters, $\sigma^2_{\epsilon_j}$ and λ. $\sigma^2_{\epsilon_j}$ is the noise power, and λ is the length parameter that determines how fast the correlation between two points decreases as the distance between the points increases. One way to find the best hyperparameter is to use cross-validation. A random sampling of training data to construct training and validation sets turned out to be inappropriate, however, since randomly sampled training and test data points are too close to each other. In this case, the obtained length parameter tends to be too small because there always exist a nearby training point to the test point. The situation is opposite to a conventional cross-validation setup, where homogeneity between training and validation data is desirable. We divided the training data into spatially disjoint cross-validation sets, and searched for λ that provides highest classification accuracies. As a result, we set $\lambda = 530$.

$\sigma^2_{\epsilon_j}$ can be interpreted as deviations from the mean function, and we estimated the deviation by calculating localized means for isolated patches as shown in Figure 2. Each patch is identified by connected component analysis. $\sigma^2_{\epsilon_j}$ is calculated as averaged squared differences of spectral values from the localized means.

3.3. Spatially Localized Priors

So far we have only considered spatial modeling of the mean vectors. In hyperspectral images, however, prior probability of each class also varies spatially[10]. We applied self-training [11], a method of semi-supervised learning, to estimated localized prior probabilities, since most of the map is unlabeled. The entire map is divided into equally sized tiles, and a classifier is applied to each tile with global priors that are estimated from the training set. For GP-ML classification, we estimate the mean vector for the center of the tile using the proposed algorithm and assume that this mean vector is common for all pixels in the tile. Classified results for each tile are then used to estimate localized prior probability distribution $P_i(\mathbf{s})$. Using the localized priors and class-conditional distributions, we run maximum-likelihood (ML) classifiers once more for each point of interest to obtain the final class label.

4. EXPERIMENTS AND RESULTS

A Hyperion hyperspectral image taken from Okavango Delta, Botswana in May 2001 is used for experiments.The acquired data originally consisted of 242 bands, but only 145 bands are used after preprocessing. The area used for experiments has 1476 × 256 pixels with 30m spatial resolution. We used two spatially disjoint class maps from the same geographical region, and there are 9 classes in total. Best-bases feature extraction algorithm [12] is used to aggregate highly correlated adjacent bands, which is beneficial for a spatially independent band model. The number of bands for best bases algorithm is 40, as determined by cross-validation. Fisher's feature extractor is also applied after best-bases extraction. The first class map is used as a training set, from which we obtain global model $p(\mathbf{x}|y_i)$ and spatially adaptive model $p(\mathbf{x}(\mathbf{s})|y_i)$, and the second map is used as a test set. For prior estimation, we used a 32 × 32 tile size. Table 1 shows classification accuracies from ML and GP-ML classifiers, before and after localized estimation of prior probabilities. Our baseline of 86.68% is the ML result without localized priors. As can be seen in the table, our algorithm with Gaussian process model and localized prior shows the best result, achieving 92.19% accuracy for a nine-class problem.

More detailed multi-class classification results are shown in Figure 3. GP-ML results are better than ML results for most classes, except for classes 4 and 8. Most probable reason for this phenomenon is that classes 4 and 8 have test points located relatively far away from the training points. The uncertainty of a Gaussian process regression model increases as the distance between training and test points increases, which makes the prediction inaccurate. A fully classified map of entire Botswana data is shown in Figure 4, where different colors indicate different land cover classes.

	ML	GP-ML
without $P_i(\mathbf{s})$	86.68%	90.59%
with $P_i(\mathbf{s})$	89.40%	**92.19%**

Table 1. Classification results from ML and GP-ML algorithms, before and after spatially localized priors

5. CONCLUSION

We have proposed a novel method for classification of hyperspectral data with a spatially adaptive approach that models class-conditional distributions by Gaussian processes, and estimates spatially localized prior probabilities with semi-supervised learning. Experimental results show that the proposed method shows significant improvements over the baseline algorithm where no spatial information is considered.

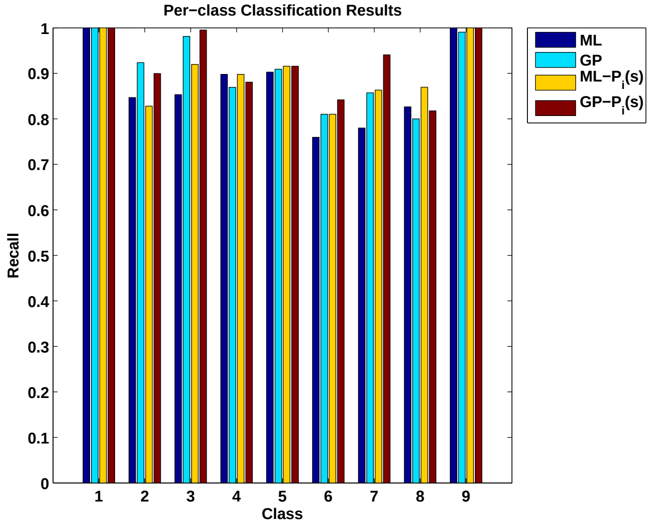

Fig. 3. Classification result for each class

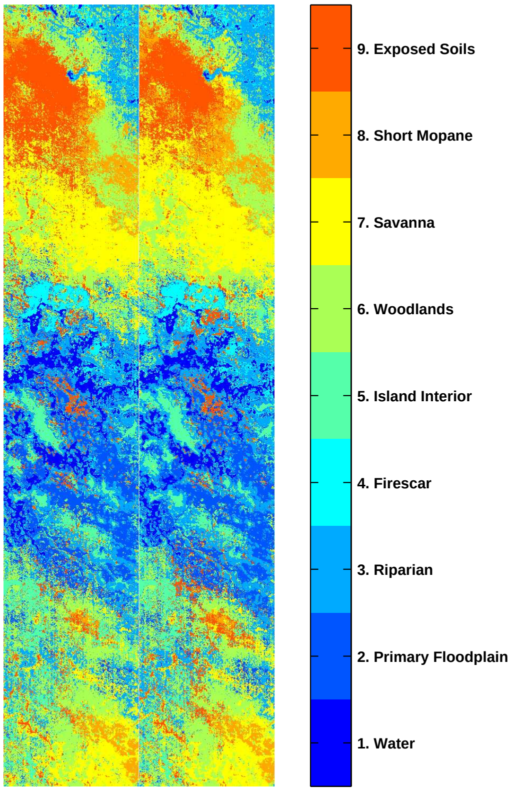

Fig. 4. Fully classified map with color-coded class labels: ML with $P_i(\mathrm{s})$ *(left)* and GP-ML with $P_i(\mathrm{s})$ *(right)*

6. REFERENCES

[1] D. Landgrebe, "Hyperspectral image data analysis," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 17–28, Jan 2002.

[2] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic Geography*, vol. 46, pp. 234–240, 1970.

[3] Andreas Krause and Carlos 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, 2007, pp. 449–456, ACM.

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

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

[6] S. Rajan, J. Ghosh, and M. M. Crawford, "Exploiting class hierarchies for knowledge transfer in hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 11, pp. 3408–3417, 2006.

[7] Yangchi Chen, M.M. Crawford, and J. Ghosh, "Knowledge based stacking of hyperspectral data for land cover classification," *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on*, pp. 316–322, 1 2007-April 5 2007.

[8] Daniel A. Griffith, "Modeling spatial dependence in high spatial resolution hyperspectral data sets," *Journal of Geographical Systems*, vol. 4, no. 1, pp. 43–51, 2002.

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

[10] 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.

[11] Xiaojin Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[12] S. Kumar, J. Ghosh, and M. M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 39, no. 7, pp. 1368–1379, 2001.