Semisupervised Learning of Hyperspectral Data With Unknown Land-Cover Classes

Goo Jun and Joydeep Ghosh, *Fellow, IEEE*

***Abstract*—Both supervised and semisupervised algorithms for hyperspectral data analysis typically assume that all unlabeled data belong to the same set of land-cover classes that is represented by labeled data. This is not true in general, however, since there may be new classes in the unexplored regions within an image or in areas that are geographically near but topographically distinct. This problem is more likely to occur when one attempts to build classifiers that cover wider areas; such classifiers also need to address spatial variations in acquired spectral signatures if they are to be accurate and robust. This paper presents a semisupervised spatially adaptive mixture model (SESSAMM) to identify land covers from hyperspectral images in the presence of previously unknown land-cover classes and spatial variation of spectral responses. SESSAMM uses a nonparametric Bayesian framework to apply spatially adaptive mechanisms to the mixture model with (potentially) infinitely many components. In this method, each component in the mixture has spatially adapted parameters estimated by Gaussian process regression, and spatial correlations between indicator variables are also considered. The proposed SESSAMM algorithm is applied to hyperspectral data from Botswana and from the DC Mall, where some classes are present only in the unlabeled data. SESSAMM successfully differentiates unlabeled instances of previously known classes from unknown classes and provides better results than the standard Dirichlet process mixture model and other alternatives.**

***Index Terms*—Clustering, Dirichlet process mixture model (DPMM), Gaussian process, hyperspectral imaging (HSI), remote sensing, semisupervised learning.**

I. Introduction

ADVANCES in remote sensing technologies have enabled identification of land covers and land usage over large geographical areas based on analysis of spectral imagery. In particular, hyperspectral imaging provides rich spectral information for each pixel and has been widely adopted for land-cover identification. Automatic classification of hyperspectral data is essential for land-cover identification problems, as a single image may contain over a million "pixels" with hundreds of spectral bands per pixel and covers large geographical areas, which makes pixelwise manual labeling impractical. Training a classifier generally requires sufficiently many labeled examples for each land-cover class of interest. In many cases, unlabeled samples are readily available in large quantities, but only a handful of land-cover labels are available due to the cost of labeling. Consequently, several semisupervised learning algorithms have been investigated for remote-sensing applications so as to utilize both the labeled and unlabeled samples for better classification. In semisupervised learning, however, the learner is unaware of the true labels of unlabeled samples; hence, it is also possible that the classifier is misinformed by the semisupervised setup. A pioneering study on the vulnerability of semisupervised algorithms in remote sensing applications was conducted in [1]. Since then, there have been several works on exploiting semisupervision, mostly focusing on improving classification accuracy when faced with limited training data. In contrast, the key contribution of this paper is to present a novel approach that enables the semisupervised learning of hyperspectral data in the presence of possibly unknown land-cover classes, where there is not even a single example of such classes in the training data. At the same time, this approach accounts for the spatial variability of data to yield very good accuracies even with limited labeled data.

Unknown land covers are possible in remotely sensed images, as the training data usually cover only a small subset of the acquired pixels. However, semisupervised learning methods developed for remote sensing applications typically assume transductive settings, where the unlabeled data are considered to have the same components as the training data and every unlabeled instance belongs to one of the classes already known to the learner. For example, the expectation–maximization (EM) algorithm for a mixture of Gaussians works well when it is applied to the test data from only known classes [2], whose labels are only *hidden* and not truly *unknown*. Given the existence of unknown classes, mixture models with a fixed number of components obtained from the training data often may be confounded by the unlabeled data. If one assumes certain probability distributions, there are algorithms that can be used to find the number of clusters in unlabeled data. For example, the number of components in the mixture model could be estimated by a simple criterion such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), where the number of components is used as a penalty term. Parametric approaches such as AIC and BIC explicitly specify the number of components and tend to oversimplify the problem. Such methods are also affected by inaccurate initial settings and local minima in the case of high-dimensional problems with many components. Recently, nonparametric Bayesian approaches based on the Dirichlet processes have gained popularity [3]. The Dirichlet process mixture model (DPMM) eliminates the need for finding the number of components explicitly by employing a mixture model with infinitely

Manuscript received October 3, 2011; revised February 16, 2012 and March 26, 2012; accepted April 18, 2012. Date of publication May 30, 2012; date of current version December 19, 2012. This work was supported by the National Science Foundation under Grant IIS-0705815.

G. Jun is with the Biostatistics Department, University of Michigan, Ann Arbor, MI 48105 USA (e-mail: gjun@umich.edu).

J. Ghosh is with the Department of Electrical and Computer Engineering, The University of Texas, Austin, TX 78712 USA (e-mail: ghosh@ece.utexas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2012.2198654

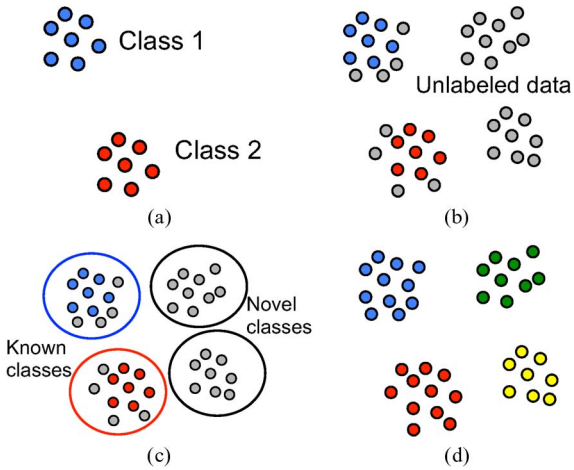

Fig. 1. Highly simplified view of the SESSAMM framework. Unknown k denotes the number of components (classes). (a) Labeled data. (b) Labeled + unlabeled. (c) Clustering with unknown k. (d) Classification.

many components, such as an infinite mixture of Gaussians [4]. The proposed semisupervised spatially adaptive mixture model (SESSAMM) takes this nonparametric-Dirichlet-process-based approach, because it provides the most flexible framework to handle mixture models with unknown number of components.

Identification of novel land-cover classes over large spatial and/or temporal extents is also challenging because the spectral response of the same land-cover class dynamically changes over space/time. For example, if the mean spectral signature of a cluster of unlabeled instances is similar but not identical to one of the known land-cover classes, it can be difficult to determine whether the difference is due to the spatial variation or because the unlabeled pixels belong to the previously unknown land-cover class. Assuming a fixed global distribution for a given class over the entire image results in larger within-class variations, which makes it more difficult to distinguish instances of the given class from similar classes.

In this paper, a novel semisupervised learning algorithm to find unknown land-cover classes from hyperspectral data is proposed by applying a spatially adaptive mixture model with (potentially) infinitely many components. This algorithm is called the SESSAMM. Fig. 1 shows a (highly simplified) view of the SESSAMM framework. First, labeled examples are assigned to their own clusters. Then, unlabeled and labeled instances are used together to find clusters using a Dirichlet-process-based clustering algorithm with spatial adaptation. The clustering results from Fig. 1(c) are used to train a supervised classifier, which classifies all unlabeled samples, as shown in Fig. 1(d). SESSAMM can employ any classifier in the framework; we used standard Gaussian maximum-likelihood (ML) and spatially adaptive Gaussian process ML (GP-ML) classifiers in this paper [5].

In SESSAMM, each mixture component employs spatially adapted parameters estimated by Gaussian process regressions. In a standard DPMM, the posterior distribution of a given instance takes a fixed form that only depends on the occupation number of each component and the concentration parameter. Such a model is too simplistic for many applications, since it does not take spatial correlation of class labels into account, i.e., cannot take advantage of the fact that neighboring pixels tend to belong to the same class. SESSAMM does not only consider spatial variations of spectral responses but also employ another Gaussian process to model spatial correlations among prior probabilities of land covers.

II. Related Work

Land-cover classification with hyperspectral data has been an active area of research in recent decades [6]–[8]. Kernel-based classification methods such as the support vector machine (SVM) have gained popularity due to the fairly high dimensionality of the data [9]–[11], where the classifier tries to find a decision boundary that maximizes separation between instances belonging to different classes in an appropriately constructed feature space. Classification algorithms are often based on a probabilistic or generative approach, such as the ML classifier which models each class with a multivariate Gaussian distribution [12]. In a generative model, the number of parameters in the model increases as the dimensionality of data increases; hence, it suffers from the curse of dimensionality and from the small-sample-size problem. To overcome such issues, a number of dimensionality reduction and feature extraction algorithms have been proposed. These include general-purpose linear feature extraction algorithms such as principal component analysis and Fisher's linear discriminant analysis (LDA) [13], as well as algorithms developed mainly for hyperspectral data analysis such as the best-base feature extraction [14], decision-boundary feature extraction (DBFE) [15], and nonparametric weighted feature extraction (NWFE) [16]. SESSAMM utilizes the best bases and Fisher's multidimensional LDA to preprocess hyperspectral data, because these feature extraction algorithms align well with SESSAMM's ML classifier with multivariate Gaussian distributions. Fisher's LDA has been shown to perform favorably with the proposed Gaussian process method as compared to DBFE and NWFE [5]; hence, the same comparison is not repeated here. We also employed the best-base feature extraction algorithm since it exploits correlations between adjacent bands and provides robust features when unlabeled data have different properties from the training data [17].

Acquiring ground reference data for a large number of examples is an expensive and time-consuming task. In contrast, unlabeled samples are easier to obtain for many problems, including land-cover classification based on remotely sensed data. Airborne or satellite images cover large geographical areas, and determining the actual land-cover type can become costly and involves much human effort, particularly in relatively inaccessible areas. Semisupervised learning refers to algorithms that exploit the unlabeled data together with the labeled data [18]. An early investigation on the usefulness of unlabeled data for hyperspectral data analysis has been done by Shahshahani and Landgrebe [1], and a plethora of semisupervised learning algorithms have been studied since then. For example, the EM algorithm can be used with the ML classification method to incorporate unlabeled samples by employing a mixture-of-Gaussians model [12]. Chi and Bruzzone presented a semisupervised SVM classification method [19]. Jia and Richards proposed a cluster-space-based algorithm that utilizes supervised and unsupervised methods together [20]. Camps-Valls *et al.* proposed a graph-based kernel method incorporating spatial information with spectral features

[21], and Tuia and Camps-Valls proposed a semisupervised method with cluster kernels [22]. Kernel-based spatiospectral methods by Camps-Valls *et al.* utilize composite kernels to encode spatiospectral information together, and our approach utilizes spatial information in preprocessing manner to separate spectral information from its spatial variation. Ratle *et al.* recently proposed semisupervised neural network classifiers [23]. Li *et al.* proposed a semisupervised segmentation algorithm for hyperspectral images that also utilizes active learning [24]. Mũnoz-Marí *et al.* proposed a one-class SVM classifier for a semisupervised setup [25]. None of these works can cater to an unknown number of novel classes in the test set without the use of an *Oracle* (as in active learning settings) and also simultaneously adapt to spatial variations in class signatures.

In remote sensing applications, it is often the case that the classifier is trained at one location and applied to other locations. Several classification algorithms have been proposed to adapt for such dynamically changing environments. Rajan *et al.* [17] proposed a knowledge-transfer framework for the classification of spatially and temporally separated hyperspectral data. Bruzzone and Persello developed a method to select spatially invariant features that provides better discrimination power when the classifier is applied to spatially distant regions [26]. There also have been studies on the active learning of hyperspectral data to minimize the required number of labeled instances to achieve the same or better classification accuracies [27], [28], and these active learning algorithms have also been tested on spatially and temporally separated data sets. Tuia *et al.* combined active learning with clustering to gain information from unlabeled regions and to discover unknown land-cover classes [29]. An active learning algorithm also exploits information from unlabeled samples, but it is different from semisupervised learning since it requires an Oracle that can produce true class labels of unlabeled instances. Chen *et al.* applied manifold techniques to analyze nonlinear variations of hyperspectral data [30], [31]. Kim *et al.* extended this manifold-based approach with multiresolution analyses [32] and proposed a spatially adaptive manifold learning algorithm for hyperspectral data analysis in the absence of sufficient labeled examples [33]. It has been shown that the Gaussian process EM (GP-EM) algorithm outperforms existing semisupervised learning algorithms for hyperspectral data [2], but it still cannot handle the existence of unknown classes.

There are algorithms that incorporate spatial information in a more direct way, such as stacking feature vectors from neighboring pixels [34]. A vector stacking approach for hyperspectral data analysis that identifies homogeneous neighborhood pixels by maximum-cut segmentation has been proposed by Chen *et al.* [35]. Image segmentation algorithms can also utilize spatial information by assuming certain levels of spatial continuity of land covers [36]–[38]. The results from these approaches largely depend on the initial segmentation results. Another possible method is majority filtering [39], where the classified map is smoothed by 2-D low-pass filters. A popular method that incorporates spatial dependences into a probabilistic model is the Markov random field model [40]–[42]. Goovaerts [43] employed a geostatistical model wherein the existence of each land-cover class is modeled by indicator kriging and combined with the spectral classification results. Kriging finds the optimal linear predictor for geospatially varying quantities [44], and the approach has been adopted in the form of Gaussian processes by machine learning researchers [45].

Recently, a classification algorithm named GP-ML has been proposed by Jun and Ghosh, where spatial variations of spectral bands are estimated by Gaussian process regressions [5]. A semisupervised version of GP-ML, i.e., GP-EM, has been also proposed by the same researchers [2], where spatial variation and semisupervised learning are addressed at the same time by employing a mixture-of-Gaussians model [46]. However, GP-EM assumes that all unlabeled samples belong to one of the known classes; hence, the performance of the algorithm may degrade significantly when there are instances from new land-cover classes. In contrast, the proposed SESSAMM algorithm employs a nonparametric Bayesian algorithm called the DPMM [3] to estimate a mixture model with unknown number of components, as in an infinite mixture of Gaussians [4]. Unlike standard DPMM, the dependent Dirichlet process (DDP) model [47] can capture covariate information between indicator variables and has been applied to modeling temporally dependent topic models [48] and spatial processes [49]. SESSAMM takes a similar approach by assuming spatially dependent posteriors on the indicator variables.

III. Background

A. DPMM

A Dirichlet distribution $(\pi_1, \ldots, \pi_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$ is a conjugate prior for a multinomial distribution and is given by

$$p(\pi_1, \ldots, \pi_k | \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{i=1}^{k} \pi_i^{\alpha_i - 1}$$

where $\pi_i \in [0, 1]$ and $\sum_{i=1}^{k} \pi_i = 1$. α_i's are parameters of the distribution, and $\Gamma(\cdot)$ is a gamma function. The Dirichlet process is a random process whose sample paths are probability distributions and whose finite dimensional distributions are Dirichlet distributions. The Dirichlet process is used to realize random draws from distributions of distributions, and it produces discrete set of output distributions, although the underlying distribution might have infinitely many possibilities. When applied to a mixture model, the Dirichlet process provides a simple way to infer a mixture model without setting the number of components *a priori*.

A mixture model with k components is defined as

$$p(\mathbf{x}|\Theta) \sim \sum_{c=1}^{k} \pi_c f_c(\mathbf{x}|\theta_c)$$

where π_c is the mixing proportion and θ_c is the parameter for the cth component. A DPMM [50] assumes a prior of symmetric Dirichlet distribution on the mixing proportion

$$(\pi_1, \ldots, \pi_k) \sim Dir\left(\frac{\alpha}{k}, \ldots, \frac{\alpha}{k}\right)$$

where *alpha* is a concentration parameter that determines how uniform the mixture is distributed. With larger value of *alpha*, the resulting mixture distribution tends to be more uniform and vice versa. Let $z_i \in \{1, \ldots, k\}$ be the membership variable of

the ith instance that represents which mixture component that the ith instance belongs to. Given fixed assignments of observed instances, the posterior distribution z_i is

$$p(z_i = c|\mathbf{z}_{-i}, \alpha) = \frac{n_c^{-i} + \alpha/k}{n + \alpha - 1}$$

where $\mathbf{z}_{-i} = \{z_j | j \neq i\}$, $n_c^{-i} = \sum_{j \neq i} \delta_{z_c,j}$, and δ is a Kronecker-delta function. Consequently, a mixture model with infinitely many components can be derived

$$\lim_{k\to\infty} p(z_i = c|\mathbf{z}_{-i}, \alpha) = \frac{n_c^{-i}}{n + \alpha - 1} \qquad \forall c, \quad n_c^{-i} > 0$$

$$\lim_{k\to\infty} \sum_c p(z_i = c|\mathbf{z}_{-i}, \alpha) = \frac{\alpha}{n + \alpha - 1} \qquad \forall c, \quad n_c^{-i} = 0.$$

n_c^{-i} is the number of elements belonging to the cth component excluding the ith instance, and $n_c^{-i} > 0$ means that the component is not empty. This formulation describes a generative model in which the prior probability of assigning an instance to an already populated component is proportional to the number of instances already belonging to the component and the probability of assigning the instance to a previously empty (novel) cluster is proportional to the concentration parameter α. The inference on the mixture model can be done by Gibbs sampling [3], as shown in Algorithm 1. The set of parameters for each component θ_c is usually estimated by defining a conjugate prior, and a special case with a mixture of Gaussians will be explained in the following section.

Algorithm 1 Outline of Gibbs sampling algorithm for a DPMM with infinitely many components

Given n instances assigned to k components,

1) For each $\mathbf{x}_i$, $1 \leq i \leq n$, do
   a) Update parameters for each component θ_c with $\mathbf{x}_i$ removed. Remove all empty components, and update k with the number of nonempty components.
   b) Calculate the likelihood of each component
   $$l_c = f(\mathbf{x}_i|\theta_c), \qquad c = 1, \ldots, k.$$
   c) Calculate the likelihood of an unpopulated component $l_{k+1} = f(\mathbf{x}_i|\theta_0)$.
   d) Calculate posteriors of z_i, $p_1, \ldots, p_{k+1}$
   $$p_c = \frac{n_c^{-i}}{n + \alpha - 1}, \qquad 1 \leq c \leq k \qquad\qquad p_{k+1} = \frac{\alpha}{n + \alpha - 1}.$$
   e) Draw $z_i \sim \text{Multi}((1/Z)p_1 l_1, \ldots, (1/Z)p_{k+1} l_{k+1})$, where $Z = \sum_{c=1}^{k+1} p_c l_c$.
   f) If $z_i = k + 1$, $k \leftarrow k + 1$.
2) Resample α and repeat.

B. Infinite Mixture of Gaussians

The DPMM could be combined with various types of distributions. Hyperspectral data can be modeled as multivariate Gaussian distributions [12]; hence, we first investigate the infinite-mixture-of-Gaussians model, where each component is modeled by a unimodal multivariate Gaussian distribution

$$f(\mathbf{x}_i|\theta_c) \sim \mathcal{N}(\boldsymbol{\mu_c}, \Sigma_c).$$

A normal-inverse-Wishart prior is employed because it is the conjugate prior for a multivariate normal distribution [51]

$$\boldsymbol{\mu_c}|\Sigma_c \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{\Sigma_c}{n_0}\right) \qquad \Sigma_c^{-1} \sim \mathcal{W}\left(m_0, \frac{\Sigma_0^{-1}}{m_0}\right).$$

$\boldsymbol{\mu}_0$, Σ_0, n_0, and m_0 are hyperparameters, where $\boldsymbol{\mu}_0$ and Σ_0 are the initial guess for the parameters and n_0 and m_0 are the pseudocounts for the mean and the covariance, respectively. These hyperparameters determine the distribution of an empty $(n_c^{-i} = 0)$ cluster, and the posterior distribution of a nonempty cluster will be pulled toward the prior distribution $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$ in proportion to the pseudocounts. $\mathcal{W}$ is a Wishart distribution. The posterior estimates of the parameters are

$$\boldsymbol{\mu}_c = \frac{1}{n_0 + n_c}(n_0\boldsymbol{\mu}_0 + n_c\bar{\boldsymbol{\mu}}_c) \tag{1}$$

$$\Sigma_c = \frac{1}{m_0 + n_c}\left(m_0\Sigma_0 + n_c\bar{\Sigma}_c + \frac{(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_c)^t}{1/n_c + 1/n_0}\right). \tag{2}$$

$\bar{\boldsymbol{\mu}}_c$ and $\bar{\Sigma}_c$ are the sample mean and covariance measured from instances assigned to the cth component. Note that $\boldsymbol{\mu}_c = \boldsymbol{\mu}_0$ and $\Sigma_c = \Sigma_0$ when $n_c = 0$, but they will move toward the sample mean and sample covariance when $n_c \gg n_0, m_0$. The likelihood of $\mathbf{x}$ from the normal-inverse-Wishart prior is a student-t, which is approximated by a moment-matched Gaussian distribution [51]

$$f(\mathbf{x}|\theta_c) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \gamma\Sigma_c), \qquad \gamma = \frac{(n_c + n_0 + 1)(n_c + m_0)}{(n_c + n_0)(n_c + m_0 - d - 1)} \tag{3}$$

where d is the dimensionality of $\mathbf{x}$, $m_0 > d + 1$.

IV. Spatially Dependent Mixtures

Although the infinite-mixture-of-Gaussians model provides a flexible representation for data with unknown number of classes, it does not incorporate spatially varying characteristics of remote sensing data. Using a single Gaussian distribution per class results in high variances, and it becomes more difficult to separate classes since there are serious overlaps between similar classes. Instead of the constant sample mean $\bar{\boldsymbol{\mu}}_c$ in (1), we employ the setup of the GP-ML algorithm presented in [5] and use a spatially adapted mean $\bar{\boldsymbol{\mu}}_c^s(\mathbf{s})$ that consists of a constant term and a spatially varying term

$$\bar{\boldsymbol{\mu}}_c^s(\mathbf{s}) = \bar{\boldsymbol{\mu}}_c + \hat{\boldsymbol{\mu}}_c(\mathbf{s}). \tag{4}$$

To obtain the spatially varying term $\hat{\boldsymbol{\mu}}_s(\mathbf{s})$, first, $\bar{\boldsymbol{\mu}}_c$ obtained from (1) is subtracted from each data point to make the data zero mean. Let X_c be an $(n_c \times d)$ matrix where each row is $(\mathbf{x}_j - \bar{\boldsymbol{\mu}}_c)^t$ for all $\mathbf{x}_j$'s with $\mathbf{z}_j = c$ and S_c be an $(n_c \times 2)$ matrix where each row is the spatial coordinate of the corresponding row in X_c. The spatially varying term $\hat{\boldsymbol{\mu}}_c(\mathbf{s})$ is obtained from a Gaussian process regression

$$\hat{\boldsymbol{\mu}}_c(\mathbf{s}) = \sigma_f^2\mathbf{k}(\mathbf{s}, S_c)\left[\sigma_f^2 K_{S_cS_c} + \sigma_\epsilon^2 I\right]^{-1} X_c$$

where $\mathbf{k}$ is a covariance vector and $K_{S_cS_c}$ is a covariance matrix. The same squared-exponential covariance function is used in the GP-ML and GP-EM algorithms. The length hyperparameter of squared-exponential covariance function is obtained by performing cross-validation using the training data, as described in [5]. In SESSAMM, the hyperparameters for GP $(\sigma_f^2, \sigma_\epsilon^2)$ are assumed to be identical across all dimensions to save computation. This simplification does not affect the result seriously when the data is prenormalized. The Gibbs sampling procedure described in Algorithm 2 requires removing and adding a single row/column from $[K_{S_cS_c} + \sigma_\epsilon^2 I]$, which can be done in $O(n_c^2)$ by using sequential updates of Cholesky decomposition, as in [52]. The adjusted sample covariance is

$$\bar{\Sigma}_c^s = \frac{1}{n_c - 1} \sum_{j:z_j=c} \left(\mathbf{x}_j - \bar{\boldsymbol{\mu}}_c^s(\mathbf{s}_j)\right)\left(\mathbf{x}_j - \bar{\boldsymbol{\mu}}_c^s(\mathbf{s}_j)\right)^t. \quad (5)$$

Using (4) and (5), (1) and (2) can be rewritten as

$$\boldsymbol{\mu}_c^s(\mathbf{s}) = \frac{1}{n_0 + n_c}\left(n_0\,\boldsymbol{\mu}_0 + n_c\,\bar{\boldsymbol{\mu}}_c^s(\mathbf{s})\right) \quad (6)$$

$$\Sigma_c^s = \frac{1}{m_0 + n_c}\left(m_0\Sigma_0 + n_c\bar{\Sigma}_c^s + \frac{(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}_c)^t}{1/n_c + 1/n_0}\right). \quad (7)$$

Consequently, the likelihood in (3) is

$$f^s(\mathbf{x}|\mathbf{s}, \theta_c) \sim \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_c^s(\mathbf{s}), \gamma\Sigma_c^s\right). \quad (8)$$

Equation (8) models the spatial variability of spectral responses only, while there also exists strong spatial correlation in the indicator variable z_i. The standard DPMM treats z_i's as independent random variables, which is not true because there are strong spatial correlations between land-cover labels, as exploited in many segmentation-based algorithms [36]–[39], or in the Markov random field model [40]–[42]. Our approach is closer to the indicator kriging approach [43], which has been successfully applied to the GP-EM [2] algorithm. For z_i, we introduce a separate Gaussian process

$$p(z_i = c|\mathbf{z}_{-i}, \mathbf{s}) \propto \sigma_z^2\mathbf{k}_z(\mathbf{s}_i, S_{-i})\left[\sigma_z^2 K_{zS_{-i}S_{-i}} + \sigma_{\epsilon_z^2} I\right]^{-1}\left(\boldsymbol{\delta}_{\mathbf{z}_{-i},c} - \frac{1}{2}\right) \quad (9)$$

where $\boldsymbol{\delta}_{\mathbf{z}_{-i},c}$ is an $(n-1)$-dimensional column vector of Kronecker-delta functions. Now, the posterior distribution of z_i is not proportional to the number of instances belonging to a certain component but depends on the proximity to the instances. A Matérn covariance function with $\nu = 3/2$ is used to calculate $\mathbf{k}_z$ and K_z, since the squared-exponential covariance function is not optimal to model abruptly changing variables such as the existence of a certain class, as discussed in [2]. Note that this distribution is no longer a posterior distribution of a Dirichlet process prior. Since our posterior is a Gaussian random process indexed by spatial coordinates, the prior for this Gaussian process is also a Gaussian process. This belongs to a family of DDPs, where each DDP is parameterized by a concentration parameter and a base random process $G_0(\mathbf{s})$ indexed by a covariate variable $\mathbf{s}$, instead of a base distribution G_0.

Algorithm 2 Outline of Gibbs sampling algorithm for SESSAMM with infinitely many components

A set of labeled data X_l with k_0 classes and a set of unlabeled data X_u are given. Initially, set $k = k_0 + 1$ by assigning labeled instances to the first k_0 components according to their class labels and assigning all unlabeled instances to the kth component. Values of indicator variables for labeled data $Z_l = \{z_i | x_i \in X_l\}$ are fixed to their known classes and not Gibbs sampled but used in the likelihood and posterior computation.

1) For each $\mathbf{x}_i \in X_u$, do
   a) Update parameters for each component with $\mathbf{x}_i$ removed. For $1 \le c \le k_0$, $\theta_c = (\boldsymbol{\mu}_c^s, \Sigma_c^s)$ from (6) and (7). For $k_0 < c \le k$, $\theta_c = (\boldsymbol{\mu}_c, \Sigma_c)$ from (1) and (2). Also update γ correspondingly.
   b) Remove all empty components, and update k with the number of nonempty components.
   c) Calculate the likelihood of each component

$$l_c = f^s(\mathbf{x}|\mathbf{s}, \theta_c) \sim \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_c^s(\mathbf{s}), \gamma\Sigma_c^s\right), \qquad 1 \le c \le k_0$$
$$l_c = f(\mathbf{x}|\theta_c) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \gamma\Sigma_c), \qquad k_0 < c \le k.$$

   d) Calculate the likelihood of an unpopulated component $l_{k+1} = f(\mathbf{x}_i|\theta_0)$.
   e) Calculate spatially adjusted posteriors of z_i from (9)

$$p_c = \frac{1}{T}q_c \equiv \frac{1}{T}\mathbf{k}(\mathbf{s}_i, S_{-i})\left[\sigma_z^2 K_{S_{-i}S_{-i}} + \sigma_{\epsilon_z}^2 I\right]^{-1} \times \left(\boldsymbol{\delta}_{\mathbf{z}_{-i},c} - \frac{1}{2}\right), \qquad 1 \le c \le k$$
$$p_{k+1} = \frac{\alpha}{n + \alpha - 1}$$

   where $T = (1 - (\alpha/n + \alpha - 1))\sum_{c=1}^{k} q_c$.
   f) Draw $z_i \sim \text{Multi}((1/Z)p_1 l_1, \ldots, (1/Z)p_{k+1}l_{k+1})$, where $Z = \sum_{c=1}^{k+1} p_c l_c$.
   g) If $z_i = k + 1$, $k \leftarrow k + 1$.
2) Resample α and repeat.

The proposed mixture model with Gaussian processes finds unlabeled instances that belong to one of the known classes effectively, but in experiments, it turned out that the algorithm is less effective for separating instances from several different unknown classes. This is mainly because the Gaussian processes adapt for instances from different classes over space and the fit is often good enough to form a single cluster. Once there is enough information from labeled instances, the fit of Gaussian processes for the cluster is stable enough to reject instances from different classes. However, in the case of clusters without any prelabeled instances, Gaussian processes for the cluster tend to adapt their mean parameters for instances from heterogeneous classes over space. To overcome this problem, a hybrid approach is taken. Spatially adjusted parameters are used only for components that have labeled instances, and a spatially invariant likelihood function with parameters in (3) is used for all other components. The outline of the Gibbs sampling procedure for the proposed SESSAMM is presented in Algorithm 2.

TABLE I
CLASS NAMES AND NUMBER OF DATA POINTS FOR BOTSWANA DATA

Class no.	Class name	# Training	# Unlabeled
1	Water	158	139
2	Primary Floodplain	228	209
3	Riparian	237	211
4	Firescar	178	176
5	Island interior	183	154
6	Woodlands	199	158
7	Savanna	162	168
8	Short mopane	124	115
9	Exposed soil	111	104

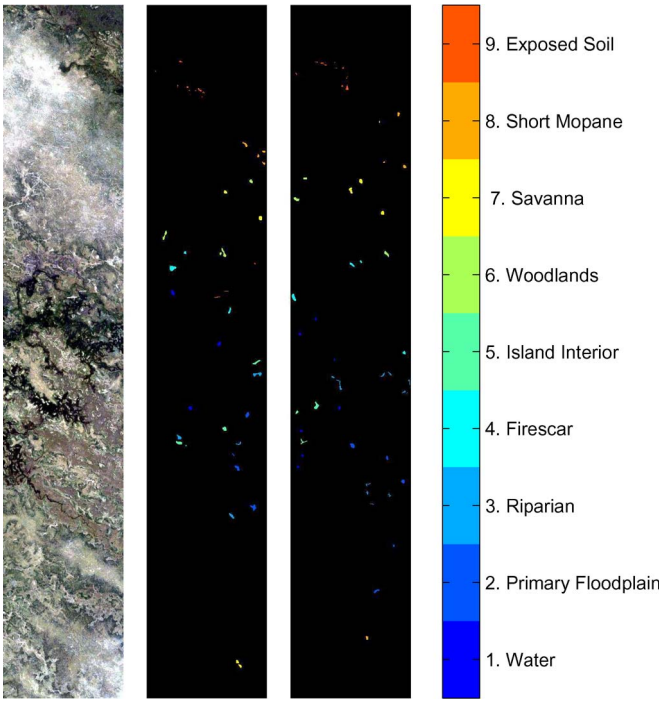

Fig. 2. Images of the nine-class Botswana data. (From left to right) Reconstructed red–green–blue image, class map of training data, and class map of test (unlabeled) data.

V. EXPERIMENTS

A. Botswana Data

The Botswana data set was obtained from the Okavango Delta by the NASA Earth Observing 1 (EO-1) satellite with the Hyperion sensor on May 31, 2001 [17], [53]. The acquired data originally consisted of 242 bands, but only 145 bands are used after removing noisy and water absorption bands. The area used for experiments has 1476 $\times$ 256 pixels with 30-m spatial resolution. The data set has spatially disjoint training and test data. The ground truth is collected using a combination of vegetation surveys, aerial photography, and a high-resolution IKONOS multispectral imagery. Table I shows the list of classes in the data with the number of training and unlabeled/test instances in each class. Fig. 2 shows the physical distribution of training and test instances in the original satellite image.

We first report results on the nine-class Botswana data set. Spatially disjoint training and test data, as shown in Table I, are used as labeled and unlabeled data sets. Randomly selected classes are removed from the training data, while the unlabeled data are used as a whole. The numbers of removed classes vary from zero to four to observe the effects of the amount of unknown classes on the clustering results. The best-base dimensionality reduction algorithm [14] is used to preprocess the data. Each band is normalized to have zero mean and unit standard deviation. The parameter α determines the prior probability of assigning an instance to an empty cluster. With larger value of alpha, chances of creating a new cluster assignment increase. Alpha is initially set to n/1000, which is rather arbitrary, but in later iterations, alpha is resampled from a noninformative prior. The Gibbs sampling is repeated 100 times for each experiment.

Table II shows the averaged clustering scores with the number of clusters obtained from DPMM and SESSAMM. Because of the nature of sampling with indefinitely many components, there are always a few instances randomly assigned to small clusters. As we are using a multivariate Gaussian distribution to model each class, small clusters that have less than 20 instances are ignored and not included in the number of clusters, as they have too few samples for stable estimation of covariance matrices. Each score is averaged over ten experiments by removing randomly selected classes from the training data. Two different metrics are used for evaluation: cluster purity and normalized mutual information (NMI) [54]. Cluster purity is a metric that indicates the proportion of cluster members that belongs to the majority class. Although the average cluster purity is an intuitive measure, it favors small clusters, and a perfect score of one is obtained when every instance is separated into singleton clusters. NMI does not favor small clusters and provides a more impartial measure. NMI is defined as

$$NMI(X,Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$$

where $H(X)$ and $H(Y)$ are the entropies of the true class distribution and the clustered results, respectively, and $I(X;Y)$ is the mutual information between them. NMI also ranges from zero to one, where a score of one means that the clustered result is identical to the ground truth. Overall, SESSAMM shows higher clustering scores than the standard DPMM in all aspects. The proposed method excels particularly in the cluster purity scores of the known classes. Compared to the standard DPMM results, there is a tendency of oversegmentation with the proposed method, where oversegmentation means that instances from a single class are sometimes divided into more than one cluster. This is mainly because pixels belonging to the same land-cover class at different spatial locations have different spectral signatures due to spatial variations, which makes it more likely for them to get assigned to different clusters. On the contrary, DPMM tends to yield undersegmented results, i.e., instances from different classes are sometimes clustered together, resulting in fewer clusters than the number of actual classes. This is mainly because DPMM prefers assigning unlabeled data to one of the already occupied clusters (i.e., known classes), as the prior probability is proportional to the number of instances belonging to the cluster. SESSAMM uses spatially adjusted priors and hence is less affected by the initial setup. In the proposed scenario of remote sensing applications, oversegmentation is more desirable than undersegmentation, since unlabeled instances from irrelevant land-cover classes could mislead the predictive model. On the other hand, a human

TABLE II
CLUSTERING RESULTS BY STANDARD DPMM WITH GAUSSIAN DISTRIBUTIONS AND BY SESSAMM WITH RANDOMLY SELECTED CLASSES REMOVED FROM THE TRAINING DATA. BOTH MEANS AND STANDARD DEVIATIONS ARE PROVIDED

# of removed classes		# clusters	Purity (known)	Purity (other)	Purity (overall)	NMI
0	DPMM	10.1 (0.316)	0.941 (0.004)	0.717 (0.073)	0.928 (0.006)	0.871 (0.006)
	SESSAMM	12.6 (0.699)	**0.990** (0.013)	**0.881** (0.040)	**0.976** (0.021)	**0.892** (0.013)
1	DPMM	9.70 (1.12)	0.891 (0.047)	0.774 (0.134)	0.877 (0.040)	0.853 (0.019)
	SESSAMM	12.9 (1.12)	**0.975** (0.033)	**0.860** (0.096)	**0.976** (0.067)	**0.882** (0.037)
2	DPMM	8.90 (0.316)	0.856 (0.054)	0.660 (0.113)	0.815 (0.038)	0.822 (0.028)
	SESSAMM	12.3 (0.949)	**0.960** (0.057)	**0.804** (0.113)	**0.908** (0.053)	**0.852** (0.032)
3	DPMM	8.40 (0.516)	0.831 (0.079)	0.656 (0.159)	0.776 (0.032)	0.794 (0.023)
	SESSAMM	15.8 (0.919)	**0.949** (0.061)	**0.866** (0.101)	**0.965** (0.067)	**0.835** (0.046)
4	DPMM	7.00 (0.667)	0.778 (0.095)	0.562 (0.131)	0.693 (0.039)	0.774 (0.028)
	SESSAMM	14.4 (2.17)	**0.910** (0.074)	**0.884** (0.088)	**0.945** (0.079)	**0.824** (0.041)

TABLE III
CLASSIFICATION ACCURACIES (IN PERCENT) OF ML AND GP-ML CLASSIFIERS WITH RANDOMLY SELECTED CLASSES REMOVED FROM THE TRAINING DATA. THE BASELINE METHOD UTILIZES LABELED SAMPLES ONLY, AND THE DPMM AND SESSAMM UTILIZE UNLABELED SAMPLES FROM CLUSTERING RESULTS. BOTH MEANS AND STANDARD DEVIATIONS ARE PROVIDED

# removed	Baseline		DPMM		SESSAMM	
	ML	GP-ML	ML	GP-ML	ML	GP-ML
0	87.2 (0.00)	91.0 (0.00)	90.3 (0.51)	85.5 (1.24)	**91.4** (0.14)	**92.8** (0.05)
1	87.9 (2.40)	91.6 (1.35)	90.5 (1.85)	85.3 (2.62)	**92.1** (1.52)	**94.5** (1.94)
2	89.3 (3.27)	91.8 (2.86)	90.6 (3.38)	86.6 (4.23)	**91.5** (3.28)	**94.2** (2.69)
3	88.8 (4.09)	91.0 (1.45)	89.6 (3.71)	85.5 (5.18)	**90.4** (3.20)	**93.4** (1.71)
4	92.6 (4.70)	91.4 (5.13)	93.0 (5.11)	90.1 (6.26)	**93.0** (5.47)	**94.7** (3.04)

TABLE IV
SESSAMM CLASSIFICATION ACCURACIES (IN PERCENT) WHEN DIFFERENT AMOUNTS OF UNLABELED SAMPLES ARE INCORPORATED FOR THE TWO-CLASS REMOVAL EXPERIMENTS

Classifier	% Unlabeled samples included					
	0%	10%	25%	50%	75%	100%
ML	89.3 (3.27)	90.2 (3.18)	90.9 (3.31)	91.3 (3.27)	91.5 (3.28)	91.5 (3.28)
GP-ML	91.8 (2.86)	93.4 (2.44)	93.6 (2.55)	94.2 (2.28)	94.0 (2.67)	94.2 (2.69)

can more easily determine that two clusters actually belong to the same class and thus correct for any oversegmentation more easily. In Table II, it can be observed that the 3 and 4 unknown class cases show high cluster purity scores than the 2 unknown class case due to oversegmentation, but the NMI scores consistently decrease with the number of unknown classes. It is remarkable that SESSAMM still shows good clustering scores even with significant numbers of classes hidden from the training data.

To evaluate how SESSAMM helps in the classification of known classes, classification accuracies for test data are reported in Table III. As we did for clustering score evaluation, the same set of random classes was removed from the training data. For the baseline ML and GP-ML results, classifiers are trained only with labeled examples. For DPMM and SESSAMM, unlabeled examples assigned to the known classes are used together with the labeled examples to train the classifier. In DPMM results, the ML classifier shows improved performances compared to the baseline ML method, but the GP-ML classifier shows inferior performances compared to the baseline GP-ML classifier. This is due to the fact that GP-ML prediction is highly dependent on the nearby examples; hence, having wrongly clustered examples in the training set significantly affects the classification results. Unlike DPMM, GP-ML after SESSAMM clustering works better than baseline GP-ML, as well as ML after SESSAMM does. We can conclude that the proposed SESSAMM framework successfully identified unlabeled examples that are helpful for classification, better than the standard DPMM clustering.

We performed another experiment to illustrate how the proposed method helps to the better identification of unexplored regions by classifying all the pixels in the image using the training data with two classes removed, the original training data, and semisupervised data clustered by SESSAMM. The classified image is provided in the online supplementary material [55]. ML classifiers are used to generate these maps to prevent extrapolation problems. In practice, the SESSAMM algorithm could be used together with any classification algorithm as it only provides clustering results. SESSAMM classification maps are generated using clustered data in addition to the seven-class training data, by assuming that an expert has identified the majority class labels of all *novel* clusters. Although novel classes are underrepresented in the SESSAMM-generated map than in the image with full training data, it is noticeable that originally hidden land-cover classes are successfully discovered. In more detailed image patches around a river, it is noticeable that the proposed method captures details of the river even better than the nine-class case. This is partly because SESSAMM benefits from the additional unlabeled data and can make better predictions for the known classes. We also tested how different amounts of unlabeled data affect classification results, and the results are shown in Table IV. From the random removal of the two-class experiments' results, SESSAMM-clustered unlabeled instances are randomly sampled at different rates. As shown in the table, the classification result improves with the increased number of unlabeled instances. It is also worth noting that GP-ML benefits more from unlabeled samples than ML, as it gains significantly with only 10% of

TABLE V
DESCRIPTION OF DC MALL DATA

Class no.	Class name	# Training	# Unlabeled
1	Roof	74	755
2	Road	54	178
3	Trail	40	89
4	Grass	88	594
5	Water	75	449
6	Tree	16	389
7	Shadow	24	73
	Total	371	2527

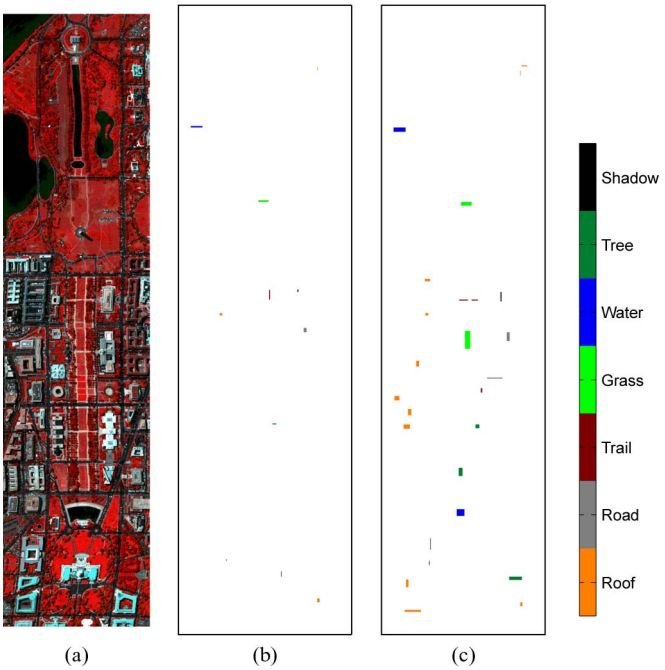

Fig. 3. (a) Simulated IR image and class maps for (b) training and (c) unlabeled data used in the experiment.

unlabeled samples while ML shows relatively smaller improvement. The same tendency has been observed in all other experiments with different numbers of classes removed.

B. DC Mall Data

An airborne hyperspectral image known as the DC Mall data [6] is used for the second set of experiments. Unlike the Botswana data, the DC Mall data contain classes from man-made objects such as building roofs, roads, and trails. As in the Botswana experiments, labeled instances are divided into spatially disjoint sets, and selected classes are removed from the training set. As shown in Table V-B, we included relatively small number of instances in the training set, compared to the unlabeled set. Fig. 3(a) shows a simulated infrared (IR) image generated from the visible and IR spectral bands of the original hyperspectral data [6]. Fig. 3(b) shows the class map of the training data used, and Fig. 3(c) shows the class map of the unlabeled data used in the SESSAMM algorithm.

Fig. 4 shows the entire images of DC Mall data classified by an ML classifier using the training data with trail and water classes removed, training data with all the classes, and semisupervised data clustered by SESSAMM. The SESSAMM-generated map originally contains oversegmented clusters, and such fragments are colored according to the major population

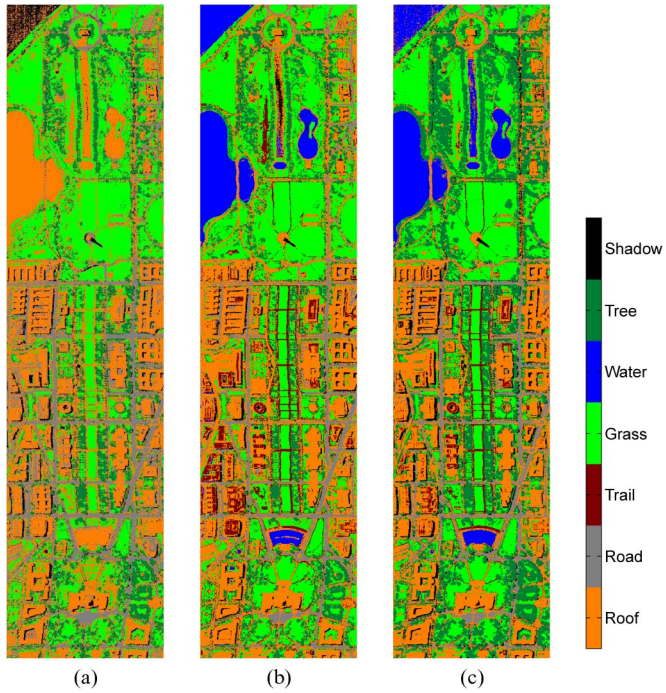

Fig. 4. Classification results from ML with five classes without water and trail, ML with seven classes, and SESSAM + ML with five classes and unlabeled data. (a) Five classes. (b) All seven classes. (c) SESSAMM.

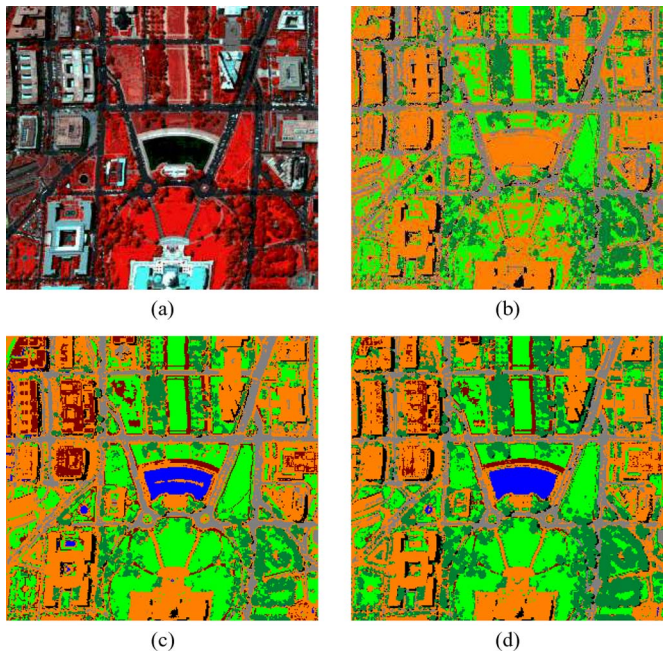

Fig. 5. Detailed classification maps of the DC Mall data around the pond. (a) Simulated IR. (b) Five classes. (c) Seven classes. (d) SESSAMM.

of the cluster for visualization purposes. Even though the water class was not included in the training data, the SESSAMM-generated map correctly identifies a pond in the lower center part of the image as water. It is noteworthy that, even with all the classes included in the training data, some part of the pond is misclassified, as shown in Fig. 4(b), which is due to the lack of nearby training data from the water class in the training data. As shown in Fig. 3(b), the training data contain water examples only in the upper left part of the

TABLE VI
DC MALL CLUSTERING RESULTS BY STANDARD DPMM WITH GAUSSIAN DISTRIBUTIONS AND BY SESSAMM WITH RANDOMLY SELECTED CLASSES REMOVED FROM THE TRAINING DATA. BOTH MEANS AND STANDARD DEVIATIONS ARE PROVIDED

# of removed classes		# clusters	Purity (known)	Purity (other)	Purity (overall)	NMI
0	DPMM	8.00 (0.00)	0.97 (0.01)	0.67 (0.17)	0.91 (0.03)	0.84 (0.02)
	SESSAMM	14.6 (0.70)	**1.00** (0.00)	**0.94** (0.07)	**0.07** (0.04)	0.82 (0.02)
1	DPMM	7.00 (0.00)	0.96 (0.03)	0.58 (0.12)	0.83 (0.08)	0.75 (0.07)
	SESSAMM	13.6 (1.26)	**0.98** (0.03)	**0.88** (0.05)	**0.94** (0.03)	**0.79** (0.02)
2	DPMM	6.00 (0.00)	0.92 (0.07)	0.51 (0.13)	0.75 (0.08)	0.70 (0.08)
	SESSAMM	12.7 (1.34)	**0.93** (0.10)	**0.93** (0.04)	**0.93** (0.05)	**0.79** (0.04)
3	DPMM	5.00 (0.00)	0.89 (0.10)	0.46 (0.19)	0.64 (0.07)	0.61 (0.09)
	SESSAMM	13.0 (1.70)	**0.98** (0.06)	**0.93** (0.05)	**0.94** (0.04)	**0.80** (0.03)

TABLE VII
CLASSIFICATION ACCURACIES (IN PERCENT) OF ML AND GP-ML CLASSIFIERS WITH RANDOMLY SELECTED CLASSES REMOVED FROM THE DC MALL DATA

# removed	Baseline		DPMM		SESSAMM	
	ML	GP-ML	ML	GP-ML	ML	GP-ML
0	83.5 (0.00)	79.4 (0.00)	84.4 (0.42)	82.5 (0.41)	**84.8** (0.56)	**82.9** (0.67)
1	82.7 (5.72)	81.1 (5.06)	84.2 (5.88)	83.6 (5.70)	**85.1** (4.72)	**83.9** (4.45)
2	84.5 (10.2)	82.8 (8.87)	90.7 (10.5)	91.6 (9.80)	**93.5** (6.12)	**93.4** (5.93)
3	86.6 (13.7)	84.7 (11.3)	95.2 (9.09)	93.9 (11.0)	**95.8** (4.66)	**94.7** (6.08)

image, and all instances in the specific patch are from the relatively deep water area. The same phenomenon is also observed at the long vertical pond in the upper center of the images. Fig. 5 zooms into the region around the pond in the maps in Fig. 4. One can note that Fig. 5(c) successfully separates originally missing classes (trail and water) from other classes.

Table VI shows the clustering scores by removing random classes from the DC Mall data, and Table VII shows the classification accuracies for nonmissing classes using training only, training plus DPMM-clustered unlabeled data, and training plus SESSAMM-clustered unlabeled data. SESSAMM-clustered results show consistently better clustering scores and better classification accuracies.

VI. CONCLUSION

The SESSAMM algorithm introduced in this paper has not only detected unlabeled instances that belong to classes that are present in the training data but also discovered novel classes when they occur in hyperspectral imagery. It achieves this feat by using a DPMM with spatial information while also accounting for spatial correlations of class labels by employing a DDP prior indexed by spatial coordinates. Experimental results show that the proposed approach provides substantially better results than the standard Dirichlet process model. Most notably, even when there is not a single example of several classes in the training data, it is able to fairly accurately discover such classes without even knowing *a priori* how many such classes there may be and with only slight oversegmentation that can be easily rectified by a human analyst.

ACKNOWLEDGMENT

The authors would like to thank M. Crawford for making available the Botswana data, for a collaboration of many years, and for the valuable comments.

REFERENCES

[1] B. Shahshahani and D. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 5, pp. 1087–1095, Sep. 1994.

[2] G. Jun and J. Ghosh, "Spatially adaptive semi-supervised learning with Gaussian processes for hyperspectral data analysis," *Statist. Anal. Data Mining*, vol. 4, no. 4, pp. 358–371, Aug. 2011.

[3] E. Sudderth, "Graphical models for visual object recognition and tracking," Ph.D. dissertation, MIT, Cambridge, U.K., 2006.

[4] C. Rasmussen, "The infinite Gaussian mixture model," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 554–560, 2000.

[5] G. Jun and J. Ghosh, "Spatially adaptive classification of land cover with remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 7, pp. 2662–2673, Jul. 2011.

[6] D. Landgrebe, "Hyperspectral image data analysis as a high dimensional signal processing problem," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 2002.

[7] E. Hestir, S. Khanna, M. Andrew, M. Santos, J. Viers, J. Greenberg, S. Rajapakse, and S. Ustin, "Identification of invasive vegetation using hyperspectral remote sensing in the California Delta ecosystem," *Remote Sens. Environ.*, vol. 112, no. 11, pp. 4034–4047, Nov. 2008.

[8] A. Plaza, J. Benediktsson, J. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, no. 1, pp. S110–S122, Sep. 2009.

[9] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[10] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005.

[11] Y. Bazi and F. Melgani, "Gaussian process approach to remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 186–197, Jan. 2010.

[12] M. Dundar and D. Landgrebe, "A model-based mixture-supervised classification approach in hyperspectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 12, pp. 2692–2699, Dec. 2002.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*., 2nd ed. New York: Wiley, 2000.

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

[15] C. Lee and D. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 4, pp. 388–400, Apr. 1993.

[16] B.-C. Kuo and D. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 5, pp. 1096–1105, May 2004.
[17] S. Rajan, J. Ghosh, and M. M. Crawford, "Exploiting class hierarchies for knowledge transfer in hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3408–3417, Nov. 2006.
[18] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, Univ. Wisconsin-Madison, Madison, WI, Tech. Rep. 1530, 2005.
[19] M. Chi and L. Bruzzone, "Semisupervised classification of hyperspectral images by SVMs optimized in the primal," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1870–1880, Jun. 2007.
[20] X. Jia and J. Richards, "Cluster-space representation for hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 3, pp. 593–598, Mar. 2002.
[21] G. Camps-Valls, T. B. Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.
[22] D. Tuia and G. Camps-Valls, "Semisupervised remote sensing image classification with cluster kernels," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 224–228, Apr. 2009.
[23] F. Ratle, G. Camps-Valls, and J. Weston, "Semisupervised neural networks for efficient hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2271–2282, May 2010.
[24] J. Li, J. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, Nov. 2010.
[25] J. Mũnoz-Marí, F. Bovolo, L. Gómez-Chova, L. Bruzzone, and G. Camp-Valls, "Semisupervised one-class support vector machines for classification of remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 8, pp. 3188–3197, Aug. 2010.
[26] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 9, pp. 3180–3191, Sep. 2009.
[27] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, Apr. 2008.
[28] G. Jun and J. Ghosh, "An efficient active learning algorithm with knowledge transfer for hyperspectral remote sensing data," in *Proc. IEEE IGARSS*, 2008, pp. I-52–I-55.
[29] D. Tuia, E. Pasolli, and W. J. Emery, "Using active learning to adapt remote sensing image classifiers," *Remote Sens. Environ.*, vol. 115, no. 9, pp. 2232–2242, 2011.
[30] Y. Chen, M. Crawford, and J. Ghosh, "Applying nonlinear manifold learning to hyperspectral data for land cover classification," in *Proc. IEEE IGARSS*, 2005, pp. 4311–4314.
[31] Y. Chen, M. M. Crawford, and J. Ghosh, "Improved nonlinear manifold learning for land cover classification via intelligent landmark selection," in *Proc. IEEE IGARSS*, 2006, pp. 545–548.
[32] W. Kim, Y. Chen, M. Crawford, J. Tilton, and J. Ghosh, "Multiresolution manifold learning for classification of hyperspectral data," in *Proc. IGARSS*, 2007, pp. 3785–3788.
[33] W. Kim, M. Crawford, and J. Ghosh, "Spatially adapted manifold learning for classification of hyperspectral imagery with insufficient labeled data," in *Proc. IEEE IGARSS*, 2008, pp. I-213–I-216.
[34] R. Haralick and K. Shanmugam, "Combined spectral and spatial processing of ERTS imagery data," *Remote Sens. Environ.*, vol. 3, no. 1, pp. 3–13, 1974.
[35] Y. Chen, M. Crawford, and J. Ghosh, "Knowledge based stacking of hyperspectral data for land cover classification," in *Proc. IEEE Symp. CIDM*, 2007, pp. 316–322.
[36] L. Jiménez, J. Rivera-Medina, E. Rodríguez-Díaz, E. Arzuaga-Cruz, and M. Ramírez-Vélez, "Integration of spatial and spectral information by means of unsupervised extraction and classification for homogenous objects applied to multispectral and hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 844–851, Apr. 2005.
[37] Y. Tarabalka, J. Benediktsson, and J. Chanussot, "Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.
[38] Y. Tarabalka, J. Benediktsson, J. Chanussot, and J. Tilton, "Multiple spectral–spatial classification approach for hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4122–4132, Nov. 2010.
[39] W. Davis and F. Peet, "A method of smoothing digital thematic maps," *Remote Sens. Environ.*, vol. 6, no. 1, pp. 45–49, 1977.
[40] Q. Jackson and D. Landgrebe, "Adaptive Bayesian contextual classification based on Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2454–2463, Nov. 2002.
[41] R. Vatsavai, S. Shekhar, and T. Burk, "An efficient spatial semi-supervised learning algorithm," *Int. J. Parallel, Emergent Distrib. Syst.*, vol. 22, no. 6, pp. 427–437, Nov. 2007.
[42] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.
[43] P. Goovaerts, "Geostatistical incorporation of spatial coordinates into supervised classification of hyperspectral data," *J. Geograph. Syst.*, vol. 4, no. 1, pp. 99–111, Apr. 2002.
[44] N. Cressie, *Statistics for Spatial Data*. New York: Wiley, 1993.
[45] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2005.
[46] V. Tresp, "Mixtures of Gaussian processes," in *Proc. NIPS*, 2001, pp. 654–660.
[47] S. MacEachern, "Dependent nonparametric processes," in *Proc. Section Bayesian Statist. Sci.*, 1999, pp. 50–55.
[48] N. Srebro and S. Roweis, "Time-varying topic models using dependent Dirichlet processes," Univ. Tornoto, UTML, Tornoto, ON, Canada, Tech. Rep. TR# 2005-003, 2005.
[49] A. Gelfand, A. Kottas, and S. MacEachern, "Bayesian nonparametric spatial modeling with Dirichlet process mixing," *J. Amer. Statist. Assoc.*, vol. 100, no. 471, pp. 1021–1035, 2005.
[50] M. D. Escobar, "Estimating normal means with a Dirichlet process prior," *J. Amer. Statist. Assoc.*, vol. 89, no. 425, pp. 268–277, Mar. 1994.
[51] A. Gelman, *Bayesian Data Analysis*. Boca Raton, FL: CRC Press, 2004.
[52] M. Seeger, "Low Rank Updates for the Cholesky Decomposition," Univ. California at Berkeley, Berkeley, CA, Tech. Rep., 2008. [Online]. Available: http://people.mmci.uni-saarland.de/mseeger/papers/cholupdate.pdf
[53] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.
[54] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining partitionings," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 583–617, 2002.
[55] G. Jun and J. Ghosh, Classified maps of Botswana and DC Mall data. [Online]. Available: http://www.ideal.ece.utexas.edu/pubs/pdf/2012/SESSAMMmap.pdf

Goo Jun received the B.S. degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1997, the M.S. degree from the University of Michigan, Ann Arbor, in 1999, and the Ph.D. degree in electrical and computer engineering from The University of Texas, Austin, in 2010.

From 1999 to 2005, he was a Research Engineer with Samsung Electronics, Suwon, Korea. He is currently a Research Fellow with the Biostatistics Department, University of Michigan.

Joydeep Ghosh (S'87–M'88–SM'02–F'06) received the B.Tech. degree from the Indian Institute of Technology, Kanpur, India, in 1983 and the Ph.D. degree from the University of Southern California, Los Angeles, in 1988.

He is currently the Schlumberger Centennial Chair Professor with the Department of Electrical and Computer Engineering, The University of Texas, Austin, where he has been with the faculty since 1988. He has published more than 250 refereed papers and 35 book chapters and coedited 20 books. Prof. Ghosh was a recipient of 14 "best paper" awards.