Self-adjusting Models for Semi-supervised Learning in Partially-observed Settings

Ferit Akova, Murat Dundar
Computer and Information Science Department
IUPUI
Indianapolis, IN, USA
Email: {ferakova, dundar}@cs.iupui.edu

Yuan Qi, Bartek Rajwa
Computer Science Department, Bindley Bioscience Center
Purdue University
West Lafayette, IN, USA
Email: alanqi@cs.purdue.edu, brajwa@purdue.edu

***Abstract*—We present a new direction for semi-supervised learning where self-adjusting generative models replace fixed ones and unlabeled data can potentially improve learning even when labeled data is only partially-observed. We model each class data by a mixture model and use a hierarchical Dirichlet process (HDP) to model observed as well as unobserved classes. We extend the standard HDP model to accommodate unlabeled samples and introduce a new sharing strategy, within the context of Gaussian mixture models, that restricts sharing with covariance matrices while leaving the mean vectors free. Our research is mainly driven by real-world applications with evolving data-generating mechanisms where obtaining a fully-observed labeled data set is impractical. We demonstrate the feasibility of the proposed approach for semi-supervised learning in two such applications.**

***Keywords*-semi-supervised learning; hierarchical dirichlet process; gaussian mixture model; partially-observed data sets; class discovery;**

I. Introduction

Despite close to two decades of active research in semi-supervised learning (SSL) there is still no consensus among researchers whether unlabeled data helps with learning. Numerous results reported over the years, with some studies showing significant improvements in classifier performance when unlabeled data is used along with labeled data, yet others presenting results [1]–[3] suggesting that semi-supervised learning is nothing but a hype, clearly indicate that the controversy surrounding semi-supervised learning will not come to an end anytime soon.

So far it has been theoretically proved that: 1. in the context of finite mixture models when the model assumption for the classifier is correct, that is, the model used to build the classifier is identical to the model that generated the data, under the additional assumption of statistical identifiability, unlabeled data alone is sufficient to identify mixture components [4]; 2. under various assumptions, classification error decreases exponentially with the number of labeled samples, and linearly with the number of unlabeled samples [5]; 3. under a zero-bias assumption, unlabeled data reduces the variance of the estimator and helps classification [6]. Although these results are strong and present the ideal conditions under which unlabeled data would be useful, the assumptions on which they are based are far from realistic for real-world data. It is now an established fact in semi-supervised learning that when there is a mismatch between approximating and true distributions, unlabeled data may actually degrade the accuracy of the classifier. Thus, it is somewhat of a paradox, to expect a distribution learned with limited labeled data to be flexible enough to accommodate a large amount of unlabeled data.

In most semi-supervised settings the limited labeled data is not only scarce but also collected without full knowledge of the underlying components of the data-generating mechanism. The main challenge that arises in the mining of real-world data sets but is often overlooked in semi-supervised learning is that the data model is not only unknown at the time of training but may also have an evolving nature that makes learning with a fixed model impractical. Under such circumstances it would be impractical to assume that labeled and unlabeled data sets come from the same distribution because certain aspects of the data-generating mechanism evident at the time the unlabeled data set was observed may not have been evident at the time the labeled data set was collected. In other words, it is natural to have a labeled dataset where the sets of classes and components are not exhaustively defined. We next present two motivating real-world applications of semi-supervised learning involving non-exhaustively defined labeled datasets.

A. Motivation

Pathogen Detection: A global surge in the number of outbreaks together with elevated concerns about biosecurity has led to an enormous interest among scientific communities and government agencies in developing label-free, i.e., reagentless, techniques for rapid identification of pathogens. The core advantage of label-free methods is their ability to quantify phenotypes for which there are no available antibodies or genetic markers. This information can be used within a traditional supervised-learning framework in which knowledge discovered from independently tested and prelabeled samples is used for training. However, the quality of training libraries is potentially limited because the sheer number of bacterial classes would not allow for practical and manageable training in a traditional supervised setting; for instance Salmonella alone has over 2400 known serovars.

Additionally, microorganisms are characterized by a high mutation rate, which means that new classes of bacteria can emerge anytime. Thus, no matter how diligently the labeled dataset is collected, the evolving nature of the problem does not allow for obtaining an exhaustively-defined labeled dataset.

Hyperspectral Data Analysis: With the recent advancements in sensor technology, remote-sensing imagery can now be collected in potentially thousands of spectral bands. This increase in spectral resolution makes it possible to differentiate land-cover types with only subtle structural differences, allowing for in-depth analysis of the scene. The widespread use of machine-learning techniques in the analysis of hyperspectral imagery is usually hindered by the lack of well-defined ground truth. Collecting ground-truth is a laborious task limited mainly by the manual labeling of the fields. The problem can get worse, especially when analyzing images of scenes that cannot be physically accessed, e.g., an enemy territory, or scenes with dynamic characteristics, e.g., urban fields. Under these circumstances, defining an exhaustive set of classes becomes impractical. The previously collected ground truth for similar scenes might allow for classification of broad land-cover types, but this comes at the expense of misclassifying fields belonging to undefined land-cover types into one of the existing types. Besides, this approach under-exploits the wealth of spectral information available in the imagery and does not allow for in-depth and high-level image analysis. In summary, the set of classes of informational value in hyperspectral image analysis is inherently non-exhaustive and like the pathogen detection problem presented above, robust analysis of hyperspectral data also requires new rigorous machine-learning approaches capable of addressing the non-exhaustiveness problem.

B. Our Approach and Contributions

Non-exhaustiveness of the labeled data set is a very realistic yet ill-defined scenario where traditional approaches to semi-supervised learning with a fixed model assumption would fail, as there is a clear mismatch between the model defined by the labeled data set and the model that generated the unlabeled data set. In this study we present a new framework for semi-supervised learning by replacing the traditional brute-force approach of fitting a fixed model onto the unlabeled data set with a new approach that can enable "data to speak for itself." We believe that our approach differs significantly from earlier work in that we relax the fixed model assumption defined by the labeled data in order to have a self-adjusting model that can evolve by dynamically adding new components or classes to better accommodate unlabeled data.

We model each class by a Gaussian mixture model (GMM) with an unknown number of components. We define a hierarchical Dirichlet process (HDP) over class distributions to dynamically model the number of components as well as classes. HDP also offers a natural framework for parameter sharing across inter- and intra-class components, practically addressing the ill-defined covariance estimation problem even for components observed with only few samples. We use a collapsed Gibbs sampler to perform inference and to estimate the posterior distribution of the component indicator variables for all samples in the labeled and unlabeled data sets. Our specific contributions in this study can be summarized as follows:

1) We propose a new framework for semi-supervised learning where unlabeled data can potentially improve learning even when the models that generated the unlabeled and labeled data sets are different.
2) We extend the concept of HDP, which allows joint learning of components across a fixed number of observed classes, to learning components of potentially infinite number of unobserved classes in addition to those of observed ones.
3) We provide a strategy for sharing the covariance matrices across different components while leaving the mean vectors free.
4) New class discovery and discovery of new components from existing classes comes as a by-product of our approach.

C. Related Work

Early work in semi-supervised learning can be broadly categorized into five areas: 1. self-training, 2. co-training, 3. graph-based methods, 4. transductive approaches, and 5. generative mixture models. We refer the reader to the survey by X. Zhu [7] for details of SSL techniques introduced in these areas. These earlier approaches in SSL assume that classes generating unlabeled data are also observed in the labeled data, which, as we discussed in the previous section, is not a realistic assumption for data sets with an evolving nature. These techniques are destined to fail when the unlabeled data set contains samples from classes not represented in the labeled data.

The work of Miller et al. [8] differs from this large body of work in SSL in that theirs is the first study to deal with the fact that the unlabeled data may originate from classes not observed in the labeled data. In this study, known and unknown classes were modeled by a mixture of expert model with learning performed by expectation-maximization [9]. The study uses minimum description length coupled with some heuristics to decide on the number of optimal mixture components, but the fully parametric nature of the proposed model does not allow for a systematic modeling of the number of mixture components. Additionally, it is not clear how the curse of dimensionality is addressed for a model that relies on the point estimates of the component parameters. It is possible that some of the mixture components present with a small number of samples in the combined labeled and unlabeled data set, in which case point estimates of the parameters will be ill-conditioned. Unlike the work of Miller et al., the proposed framework uses a non-parametric prior to model the number of components and adopts a fully Bayesian approach for model learning that eliminates the need for point estimates of the parameters, practically addressing the curse of dimensionality problem to a greater extent.

Most of what we did in this study is more closely related to the field of Bayesian non-parametrics primarily involving Dirichlet process mixture (DPM) models [10]–[12]. The DPM has been heavily studied for clustering applications over the past decade. Most of these approaches assume that all the components of the mixture model are unobserved and study inference techniques to learn these components in an unsupervised manner. Although certain aspects of these studies have been inherently useful for our study, an unsupervised approach would be most desirable in settings where the patterns and structure within the data set are completely unobserved.

Two recent studies [13], [14] that use a partially-observed DPM for class modeling are of particular interest for the current study. Both of these techniques deal with online learning problems and use a partially observed DPM to model existing and emerging classes together. The work of Dubey et al. [13] models training data by a HDP and introduces a DP model to handle incoming data. Incoming data contains samples from observed as well as unobserved classes. HDP and DP models are then coupled with the goal of identifying news articles with new topics while classifying those with older topics into one of the classes represented in the training data. In [14], motivated by a pathogen detection problem, a partially observed DPM model is coupled with a particle filtering algorithm to detect emerging classes in an online manner.

We believe that the proposed framework pioneers the approach to learning with a non-exhaustively defined labeled data set and presents a unique framework to tackle unlabeled data in partially-observed semi-supervised settings. The distinct algorithmic aspects of the proposed study involve, first, the extension of the HDP model to accommodate unlabeled data and to discover and recover new classes, and second, the fully Bayesian treatment of mixture components to allow for parameter sharing across different components that not only addresses the curse of dimensionality problem but also connects observed classes with unobserved ones through sharing of parameters.

The rest of this paper is organized as follows. In Section II-A we briefly review the hierarchical dirichlet process (HDP). In Section II-B we discuss how HDP can be extended to partially observed settings. In Section II-C we incorporate the data model and discuss a strategy for sharing the covariance matrices while leaving the mean vectors free. In Section II-D we demonstrate the feasibility of the proposed approach on an artificial dataset. In Section II-E we discuss some of the implementation details. In Section III we present results of our experiments comparing the proposed approach against several state-of-the-art supervised and semi-supervised learning techniques for the bacteria classification and hyperspectral image analysis problems. In Section IV we conclude with a brief summary of our contributions and future research directions.

II. Learning from Nonexhaustive Data

We start this section with a brief review of the Hierarchical Dirichlet Processes (HDP) [15] widely used in the machine learning literature for co-clustering multiple groups of data by enabling sharing of parameters across components. Throughout this section we use the terms *group* and *class* interchangeably. We also assume that each group data comes from a mixture model with an unknown number of components.

A. Hierarchical Dirichlet Processes

HDP extends Dirichlet Processes (DP) [11], which is mainly used in clustering and density estimation problems as a nonparametric prior defined over the number of mixture components. A DP can be considered as a distribution over distributions. HDP models each group of data in the form of a DPM model, where DPM models across different groups are connected together through a higher level DP. We use the notation $x_{ji} \in \Re^d$, $i = \{1, ..., n_j\}$, $j = \{1, ..., J\}$ to identify sample i in the group j where n_j denotes the number of samples in group j, J is the total number of groups, and θ_{ji} defines the parameters of the mixture component associated with x_{ji}. Each x_{ji} is associated with a mixture component defined by the parameter θ_{ji}, which is generated i.i.d. from a Dirichlet Process as follows:

$$\begin{array}{lcll} x_{ji}|\theta_{ji} & \sim & p(\cdot|\theta_{ji}) & \text{for each j, i} \\ \theta_{ji}|G_j & \sim & G_j & \text{for each j, i} \end{array} \tag{1}$$

where G_j's are random probability measures distributed i.i.d. according to a DP with base distribution G_0 and precision parameter α. The stick-breaking construction due to [16] suggests $G_j = \sum_{i=1}^{\infty} \beta_{ji}\delta_{\theta_{ji}}$ where $\beta_{ji} = \beta'_{ji}\prod_{l=1}^{i-1}(1-\beta'_{jl})$, $\beta'_{ji} \sim Beta(1, \alpha)$, and $\theta_{ji} \sim G_0$. The points θ_{ji} are called the *atoms* of G_j. Note that unlike continous distributions the probability of sampling the same θ_{ji} twice is not zero and proportional to β_{ji}. Thus, G_j is considered a discrete distribution.

The precision parameter, α, is the parameter that controls how much of the stick will be left for subsequent values. The smaller the α is, the larger the β'_{ji} will be, and the less of the stick will be left for subsequent values on average. Thus, α is the parameter that controls the prior probability of assigning a new sample to a new component and thus, plays a critical role in the number of components generated.

In the HDP model the base distribution G_0 is distributed according to a higher level DP with a base distribution H and parameter γ. This hierarchical model couples G_j's and allow for sharing of mixture components within and between groups. HDP model is completed as follows:

$$\begin{array}{rcll} G_j|G_0,\alpha & \sim & DP(G_0,\alpha) & \text{for each j,} \\ G_0|H,\gamma & \sim & DP(H,\gamma) & \end{array} \tag{2}$$

The generative process defined by an HDP model can be better explained by an analogy to the Chinese Restaurant Franchise (CRF) [15]. We have a restaurant franchise with a global menu of dishes shared across all restaurants. In each restaurant a certain dish is served at each occupied table, which is shared by all customers sitting in that table. The same dish can be served in other tables across multiple restaurants. The popularity of a particular dish is proportional to the number of tables serving that dish. In an arbitrary restaurant j, customer i is associated with θ_{ji} and is seated at table t_{ji}, and table t is associated with one of the K random draws from H, i.e., $\psi_{jt} \in \{\phi_1,\ldots,\phi_K\}$, which represents the global menu of dishes. A dish from the global menu served at table t in restaurant j is denoted by the indicator variable k_{jt}. In the HDP model the parameter γ controls the prior probability of serving a new dish at a new table.

In this model, restaurants correspond to classes, each table in a restaurant corresponds to a mixture component in the mixture model, and each dish in the menu corresponds to a unique set of parameters shared by one or more components.

The conditional distributions for t_{ji} and k_{jt} are obtained by integrating out G_j and G_0, respectively.

$$t_{ji}|t_{j1},\ldots,t_{j,i-1},\alpha \quad \sim \quad \tfrac{\alpha}{n_j+\alpha}\delta_{t^{new}} + \textstyle\sum_{t=1}^{m_{j\cdot}} \tfrac{n_{jt}}{n_j+\alpha}\delta_t \tag{3}$$

where $m_j.$ is the number of tables in restaurant j and n_{jt} is the number of customers at table t in restaurant j. According to this conditional distribution θ_{ji} inherits one of the existing ψ_{jt} with probability $\frac{n_{jt}}{n_j+\alpha}$ or $\psi_{j,m_j.+1}$, i.e., a new table, with probability $\frac{\alpha}{n_j+\alpha}$. Similarly,

$$k_{jt}|k_{j1},\ldots,k_{j,t-1},\gamma \quad \sim \quad \tfrac{\gamma}{m_{\cdot\cdot}+\gamma}\delta_{k^{new}} + \textstyle\sum_{k=1}^{K} \tfrac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}\delta_k \tag{4}$$

where $m_{.k}$ is the number of tables across all restaurants serving dish ϕ_k and $m_{..}$ is the total number of tables across all restaurants. According to this conditional distribution ψ_{jt} is equal to one of the ϕ_k with a probability $\frac{m_{.k}}{m_{..}+\gamma}$ or ϕ_{K+1}, i.e., a new dish, with probability $\frac{\gamma}{m_{..}+\gamma}$.

Inference in the described CRF setting can be performed using a Gibbs sampler by iteratively sampling the variables $\boldsymbol{t} = \left\{\{t_{ji}\}_{i=1}^{n_j}\right\}_{j=1}^{J}$, $\boldsymbol{k} = \left\{\{k_{jt}\}_{t=1}^{m_{j\cdot}}\right\}_{j=1}^{J}$, and $\boldsymbol{\phi} = \{\phi_k\}_{k=1}^{K}$ given the state of all other variables. The conditional distributions for t_{ji} is:

$$p(t_{ji}=t|\boldsymbol{t}\backslash t_{ji},\boldsymbol{k},\boldsymbol{\phi},\boldsymbol{x}) \propto \\ \begin{cases} \alpha p(x_{ji}) & \text{for } t = m_{j\cdot}+1 \\ n_{jt}^{-i} p(x_{ji}|\phi_{k_{jt}}) & \text{for } t \in \{1,\ldots,m_{j\cdot}\} \end{cases} \tag{5}$$

The conditional distributions for k_{jt} is:

$$p(k_{jt}=k|\boldsymbol{t},\boldsymbol{k}\backslash k_{jt},\boldsymbol{\phi},\boldsymbol{x}) \propto \\ \begin{cases} \gamma \prod_{i:t_{ji}=t} p(x_{ji}) & \text{for } k = K+1 \\ m_{.k}^{-jt} \prod_{i:t_{ji}=t} p(x_{ji}|\phi_k) & \text{for } k \in \{1,\ldots,K\} \end{cases} \tag{6}$$

In the above conditional distributions n_{jt}^{-i} is the number of customers sitting at table t in restaurant j excluding the customer i, $m_{.k}^{-jt}$ is the number of tables sharing the same dish ϕ_k excluding the table t in the restaurant j. The conditional distribution for $\boldsymbol{\phi}$ is omitted as we choose a conjugate pair of H and $p(\cdot|\phi)$ in this study, which allows us to integrate out ϕ analytically to obtain a collapsed version of the Gibbs sampler.

B. Partially-observed HDP model for semi-supervised learning

In this section we extend the HDP model to semi-supervised learning in partially-observed settings. We model each class by a Gaussian mixture model (GMM) with an unknown number of components. We introduce the notion of observed and unobserved classes/subclasses to distinguish classes/subclasses represented in the labeled data set from those not represented. Each subclass is represented by a single component in the corresponding GMM model. Thus, we use subclasses and components interchangeably in the rest of the paper.

The labeled data set is non-exhaustively defined because the set of classes and the set of components for some or all of the classes are not complete, i.e., partially observed. The class labels for samples in the labeled data set are known, whereas component labels are not. The unlabeled data set may contain samples from classes and subclasses not represented in the labeled data set. However, neither the class labels nor the component labels are known for samples in the unlabeled data set. The number of components in each class and the total number of classes are also not known.

In the partially-observed setting the learning problem includes the following two tasks: (i) inferring the component membership of labeled samples and (ii) inferring both the group and component membership of unlabeled samples. Unlike labeled samples which are known to originate from observed classes, unlabeled samples can originate from observed as well as unobserved classes. Each class in the proposed SSL framework corresponds to a separate restaurant in the CRF analogy. To relate this partially observed setting to the CRF metaphor each unlabeled sample can be considered as an *undecided* customer who has not yet decided which restaurant to go. These customers can go to one of the existing restaurants in the franchise but may as well choose a new restaurant, which is treated as a new restaurant with a single table in the proposed framework. Labeled samples represent customers who already arrived at one of the existing restaurants and waiting to be seated. These customers can be seated using the same Gibbs sampler scheme presented in the previous section after accounting for the presence of undecided customers who eventually choose to go to the same restaurant. In short, *decided* customers sit at existing or new tables in existing restaurants only, whereas *undecided* customers can seat at new tables in new restaurants in addition to existing or new tables in existing restaurants. Before we move on to describing the details of our approach for extending the HDP framework for semi-supervised learning in a partially-observed setting we introduce new notation to distinguish between labeled and unlabeled samples.

We use $\boldsymbol{x} = \left\{\{x_{ji}\}_{i=1}^{n_j}\right\}_{j=1}^{J}$ and $\boldsymbol{t} = \left\{\{t_{ji}\}_{i=1}^{n_j}\right\}_{j=1}^{J}$ to denote samples and component indicator variables, respectively, for the labeled data. For the same variables in the unlabeled data, we use $\tilde{\boldsymbol{x}} = \{\tilde{x}_i\}_{i=1}^{n_u}$ and $\tilde{\boldsymbol{t}} = \{\tilde{t}_i\}_{i=1}^{n_u}$. For the unlabeled data we also introduce $\tilde{\boldsymbol{y}} = \{\tilde{y}_i\}_{i=1}^{n_u}$ to denote the unknown class indicator variables, where $\tilde{y}_i \in \{1, \ldots, J + \bar{J}\}$, $\bar{J}$ is the number of newly created classes after observing the unlabeled data. Finally we use $\boldsymbol{k} = \left\{\{k_{jt}\}_{t=1}^{m_{j.}}\right\}_{j=1}^{J}$ and $\tilde{\boldsymbol{k}} = \left\{\tilde{k}_j\right\}_{j=1}^{\bar{J}}$ to define indicator variables for the unique parameters shared across observed and newly created classes, respectively.

The part of the Gibbs sampler for inferring the component membership of labeled samples involve evaluating the following conditional distributions iteratively given the state of all other variables.

The conditional distribution for t_{ji} for a labeled sample is:

$$p(t_{ji} = t | \boldsymbol{t} \backslash t_{ji}, \boldsymbol{k}, \boldsymbol{\phi}, \boldsymbol{x}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{t}}) \propto \\ \begin{cases} \alpha\, p(x_{ji}) & \text{for } t = m_{j.} + 1 \\ (n_{jt}^{-i} + \tilde{n}_{jt}) p(x_{ji} | \phi_{k_{jt}}) & \text{for } t \in \{1, \ldots, m_{j.}\} \end{cases} \tag{7}$$

where $\tilde{n}_{jt}$ is the number of unlabeled samples assigned to component t in class j.

Unlike a labeled sample, which is either assigned to one of the existing components associated with its class of origin or to a new component generated for that class, an unlabeled sample can be assigned to any of the existing components across all classes or to a new component generated for a new class. In this framework each new class will inherently have one component. The fact that true labels of unlabeled samples are not known makes it impossible to readily associate new components with existing ones.

The conditional distribution for $\tilde{t}_i$ for an unlabeled sample is:

$$p(\tilde{t}_i = t | \boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\phi}, \boldsymbol{x}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{t}} \backslash \tilde{t}_i, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{k}}) \propto \\ \begin{cases} \alpha\, p(\tilde{x}_i) & \text{for } t = 1 \\ & j = J + \bar{J} + 1 \\ (n_{jt} + \tilde{n}_{jt}^{-i}) p(\tilde{x}_i | \phi_{k_{jt}}) & \text{for } t \in \{1, \ldots, m_{j.}\} \\ & j \in \{1, \ldots, J + \bar{J}\} \end{cases} \tag{8}$$

Next, we discuss the part of the Gibbs sampler for inferring the indicator variables of unique parameters for components of existing and new classes.

A component in an existing class may contain both labeled and unlabeled samples. Thus, the conditional distribution for k_{jt} is:

$$p(k_{jt} = k | \boldsymbol{t}, \boldsymbol{k} \backslash k_{jt}, \boldsymbol{\phi}, \boldsymbol{x}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{t}}, \tilde{\boldsymbol{k}}) \propto \\ \begin{cases} \gamma \prod_{i:t_{ji}=t} p(x_{ji}) \prod_{i:\tilde{t}_i = t \wedge \tilde{y}_i = j} p(\tilde{x}_i) \\ \quad \text{for } k = K + 1 \\ m_{.k}^{-jt} \prod_{i:t_{ji}=t} p(x_{ji} | \phi_k) \prod_{i:\tilde{t}_i = t \wedge \tilde{y}_i = j} p(\tilde{x}_i | \phi_k) \\ \quad \text{for } k \in \{1, \ldots, K\} \end{cases} \tag{9}$$

On the other hand a component in a new class contains only unlabeled samples. Thus, the conditional distribution for $\tilde{k}_j$ is:

$$p(\tilde{k}_j = k | \boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\phi}, \boldsymbol{x}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{t}}, \tilde{\boldsymbol{k}} \backslash \tilde{k}_j) \propto \\ \begin{cases} \gamma \prod_{i:\tilde{y}_i = j} p(\tilde{x}_i) \\ \quad \text{for } k = K + 1 \\ m_{.k}^{-j} \prod_{i:\tilde{y}_i = j} p(\tilde{x}_i | \phi_k) \\ \quad \text{for } k \in \{1, \ldots, K\} \end{cases} \tag{10}$$

Finally, the class indicator variables $\tilde{\boldsymbol{y}}$ for unlabeled samples can be obtained from $\tilde{\boldsymbol{t}}$. If an unlabeled sample is assigned to a new component this will indicate a new class and thus $\tilde{y}_i = J + \bar{J} + 1$. If an unlabeled sample is assigned to one of the existing components associated with class j then $\tilde{y}_i = j$. Note that class j can be one of the classes represented in the labeled data set as well as one of the classes previously associated with unlabeled samples, i.e., $j \in \{1, \ldots, J + \bar{J}\}$. Each sweep of the Gibbs sampler also involves sampling γ and α values using the technique described in [10].

This completes our discussion for learning with labeled and unlabeled data sets with an HDP model in a partially-observed setting. Next, we will present the data model used in this study and discuss a strategy for sharing the covariance matrices of mixture components while leaving their mean vectors free.

C. Parameter Sharing in a Gaussian Mixture Model

We model each class by a mixture model with each component data distributed according to a Gaussian distribution with mean vector μ_{jt} and a covariance matrix Σ_{jt}, i.e., $\psi_{jt} = \{\mu_{jt}, \Sigma_{jt}\}$. For the base distribution H, from which the unique component parameters ϕ_k's are sampled, we define a conjugate prior:

$$H = p(\mu, \Sigma) = \underbrace{\mathcal{N}\left(\mu|\mu_0, \frac{\Sigma}{\kappa}\right)}_{p(\mu|\Sigma)} \times \underbrace{W^{-1}(\Sigma|\Sigma_0, m)}_{p(\Sigma)} \quad (11)$$

where μ_0 is the prior mean and κ is a scaling constant that controls the deviation of the mean vectors of mixture components from the prior mean. The smaller the κ is, the larger the scattering between the components will be. The parameter Σ_0 is a positive definite matrix that encodes our prior belief about the expected Σ. The parameter m is a scalar that is negatively correlated with the degrees of freedom. In other words the larger the m is the less Σ will deviate from Σ_0 and vice versa. The parameters $(\Sigma_0, m, \mu_0, \kappa)$ are estimated using labeled samples in the same way described in our earlier work [14].

To evaluate the Gibbs sampler introduced in the previous section we need the conditional distribution $p(x|\psi_{jt})$ and the marginal distribution $p(x)$. Since ψ_{jt} are not known they can be replaced with the class conditional predictive distributions $p(x|D_{jt})$, where D_{jt} denotes the subset of samples belonging to component t in class j. This collapsed version of the Gibbs sampler eliminates the need for point estimates of ψ_{jt} and reduces the state space of the sampler leading to faster convergence to the equilibrium distribution [17]. The marginal distribution can be obtained from $p(x|D_{jt})$ by setting D_{jt} an empty set. For the multivariate Gaussian data the sample mean $\bar{x}$ and the sample covariance matrix S are sufficient statistics and therefore we can write $p(x|D_{jt}) = p(x|\bar{x}_{jt}, S_{jt})$.

To evaluate $p(x|D_{jt})$ for a given x requires evaluating the following integral with respect to $\psi_{jt} = \{\mu_{jt}, \Sigma_{jt}\}$.

$$p(x|D_{jt}) = \int p(x|\psi_{jt})p(\psi_{jt}|D_{jt})\partial\psi_{jt} \quad (12)$$

If parameter sharing across different components were not allowed, evaluating the above integral analytically would yield a multivariate Student-t distribution with the following parameters [18].

Location vector:

$$\hat{\mu}_{jt} = \frac{n_{jt}\bar{x}_{jt} + \kappa\mu_0}{n_{jt} + \kappa}$$

Scale matrix:

$$\hat{\Sigma}_{jt} = \frac{\Sigma_0 + (n_{jt}-1)S_{jt} + \frac{n_{jt}\kappa}{n_{jt}+\kappa}(\bar{x}_{jt} - \mu_0)(\bar{x}_{jt} - \mu_0)^T}{\frac{(\kappa+n_{jt})\,v}{(\kappa+n_{jt}+1)}} \quad (13)$$

Degrees of freedom:

$$v_{jt} = m + n_{jt} - d + 1$$

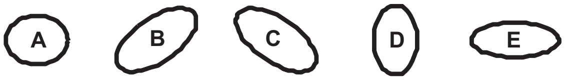


Figure 1. Templates of covariance matrices used for the illustrative example

However, in the proposed framework the clustering property of the HDP model allows multiple components to inherit one of the distinct parameters in $\boldsymbol{\phi}$. Thus, instead of integrating out ψ_{jt} as in (12), sharing property of the HDP model requires that we integrate out ϕ_k in the predictive distribution. Let $D_{.k}$ be the samples of all components sharing parameter ϕ_k then the predictive distribution $p(x|D_{.k})$ can be obtained by evaluating the following integral.

$$p(x|D_{.k}) = \int p(x|\phi_k)p(\phi_k|D_{.k})\partial\phi_k \quad (14)$$

If ϕ_k contains both the mean vector and the covariance matrix then this would imply sharing the same mean vector and the covariance matrix across multiple components. This would mean fitting each component by the same Gaussian distribution, which would not make sense as components sharing the same parameters would no longer be identifiable. To tackle this problem we adopt a strategy, where sharing is limited with the covariance matrices only. Thus, if we set $\phi_k = \{\Sigma_k\}$ and evaluate the integral in (14) analytically we obtain the predictive distribution as a multivariate Student-t distribution with the same location vector as previously but with the scale matrix and degrees of freedom updated as follows.

Scale matrix:

$$\hat{\Sigma}_{jt} = \frac{\Sigma_0 + \sum_{jt:k_{jt}=k}(n_{jt}-1)S_{jt} + \frac{n_{jt}\kappa}{n_{jt}+\kappa}(\bar{x}_{jt}-\mu_0)(\bar{x}_{jt}-\mu_0)^T}{\frac{(\kappa+n_{jt})\,v}{(\kappa+n_{jt}+1)}} \quad (15)$$

Degrees of freedom:

$$v_{jt} = m + \sum_{jt:k_{jt}=k}(n_{jt} - 1) - d + 2$$

where the summation terms are over all components sharing the same covariance matrix.

Next, we present an illustrative example demonstrating the proposed approach discovering and recovering new classes and new components of observed classes.

D. Illustration of the Proposed Approach

For this illustration we generated three classes, each as a mixture of three Gaussian components. The covariance matrices for individual components are randomly drawn from a set of five different templates of covariance matrices, each with a different shape and orientation (Figure 1). The mean vectors of the classes are equidistantly placed along the periphery of a circle centered at the origin with radius 7. Similarly, the component means are arbitrarily chosen along a circle with radius 1, centered at the corresponding class means.

We generated 110 samples from each component for a total of 330 samples for each class. We randomly selected 10 samples from each component as labeled data and used the remaining 100 samples from that component as unlabeled data. In order to produce a partially-observed labeled data set in terms of both the number of classes and the number of components for an observed class, we considered all components of a class and one component of a second class as unobserved and discarded all their labeled samples, leaving only unlabeled samples only these components.

The purpose of this illustration is three fold. First, we show that the proposed HDP model, which uses unlabeled and labeled data together, can more accurately recover the underlying distributions of the observed classes compared to the version that uses only labeled data. Second, we demonstrate that the proposed self-adjusting model can successfully discover and recover the underlying distributions of classes/subclasses that exist in the unlabeled data but are unobserved in the labeled data, whereas classical approaches that deal with samples of unrepresented classes/subclasses by assigning reduced weight to them can neither discover unobserved classes nor accurately model observed classes. Third, we illustrate the sharing aspect of the proposed approach by first identifying the types of the covariance matrices of the recovered distributions and then comparing them against the true types of the covariance matrices used to generate data from each subclass. We show that with the proposed approach the labels of the covariance matrices shared among recovered subclass distributions perfectly match the labels of those shared among true subclass distributions.

Figure 2a shows true subclass distributions for all nine subclasses. The observed subclasses, i.e., those that are represented in the labeled data set, are shown by solid lines and unobserved ones by dashed lines. The ellipses correspond to the distributions of the subclasses that are at most three standard deviations away from the mean.

Figure 2b shows the distributions of the five observed subclasses recovered by the version of the HDP model that uses only labeled data. Note that the recovered distributions deviate from the true subclass distributions. Additionally three of the five recovered subclass distributions share different types of covariance matrices than those used in the true subclass distributions.

Figure 2c shows the impact of unlabeled data over the recovered subclass distributions when unlabeled data contain samples from classes/subclasses unobserved in the labeled data and a fixed model is used to accommodate unlabeled data. These results are obtained using the technique introduced in [19], which assigns reduced weight to unlabeled samples as determined by their posterior probabilities. Note that, since unlabeled data from unobserved subclasses dominate labeled data from observed classes, the recovered distributions for observed classes significantly deviate from true subclass distributions.

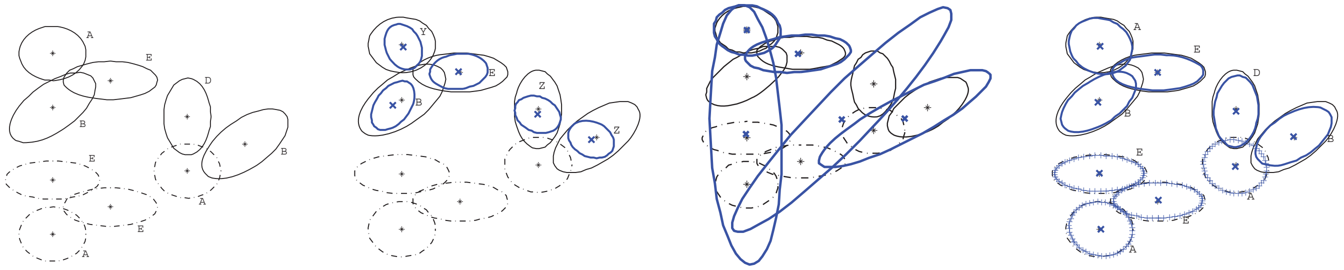
Figure 2d shows the results of the proposed approach. Both observed and unobserved subclass distributions are almost perfectly recovered. The sharing of the covariance matrices among recovered subclass distributions matches the sharing of covariance matrices among true subclass distributions. Labels for the covariance matrices of recovered distributions are also correctly identified.

E. Implementation of the Proposed Approach

We end this section by briefly discussing some of the implementation details involving the Gibbs sampler presented in Section II-B. We initialize the HDP model by generating a component for each observed class and assigning a random sample from that class to this component. During each sweep of the Gibbs sampler, all data samples are assigned to one of the existing components or to a new component using equations (7) and (8) for labeled and unlabeled samples, respectively. This is followed by sampling the parameters of the components based on the most current assignment of the samples. For components associated with observed classes the equation (9) is used, for those associated with unobserved classes the equation (10) is used. Both labeled and unlabeled samples can generate new components but unlike a component generated for an unlabeled sample, which is assigned to a new class, a new component generated for a labeled sample is readily assigned to the observed class the labeled sample belongs to.

When a sample is assigned to an existing component the mean vector of the corresponding component and the covariance matrices of all components associated with the same ϕ are updated. If an unlabeled sample ends up at a new table then we introduce a new class and tentatively label the sample with that class until the next iteration and process remaining unlabeled samples by taking the new table into account as well. Unlike an observed class, which is fixed by definition, a new class can be removed when no samples are left in the component associated with that class. A component associated with an observed class may contain both labeled and unlabeled samples at a given sweep but during later sweeps the labeled samples may move to other components leaving only unlabeled ones in that component. In this case we reassign that component to a new class.

Finally, as mentioned before, we used the formulation in [10] to sample the precision parameters α and γ of the HDP model for the Gibbs sampler. This formulation requires defining Gamma priors with shape parameters $(a0, b0)$ and $(a1, b1)$ over α and γ, respectively. The posterior distribution for α is conditioned on the total number of samples N and the total number of existing components $m_{..}$ in the current



(a) (b) (c) (d)

Figure 2. Illustration of the proposed algorithm with an artificial dataset. Solid and dashed black contours indicate observed and unobserved subclasses, respectively. Solid blue contours indicate recovered versions of observed subclasses whereas blue contours plotted with the plus sign indicate the recovered versions of unobserved subclasses. Letters denote the labels of the covariance matrices. The star and cross signs show the location of the true and predicted mean vectors of subclasses, respectively. (a) True subclass distributions. (b) Distributions recovered by the standard HDP model using only labeled data set. (c) Distributions recovered by a fixed model that assigns full weight to labeled samples and reduced weight to unlabeled samples, using both labeled and unlabeled data sets. (d) Distributions recovered by the proposed self-adjusting model using both labeled and unlabeled data sets.

iteration. Similarly, the posterior for γ is conditioned on $m_{\cdot\cdot}$ and the number of existing unique parameters in the current iteration K. While experimenting with the shape parameters of the Gamma priors, we observed that as $m_{\cdot\cdot}$ increases it dominates the effect of $a0$ in the posterior and the expected value of the Gamma posterior tends to increase regardless of $a0$. Regarding the second parameter of the posterior, as N is fixed, a large value for $b0$ is necessary to balance the effect of the first parameter on the expected value. A similar argument can be made for γ based on the values of K and T. As a result we set $a0$ and $a1$ to one and coarsely tuned $b0$ and $b1$ values.

III. Experiments

A. Evaluated Classifiers

We considered three supervised learning methods as baseline techniques, where only the labeled training samples is used for learning the classifiers. The first one is a Naive Bayes classifier (SL-NB). The second one is a maximum-likelihood classifier with each class modeled by a single Gaussian (SL-ML). The third one is a maximum-likelihood classifier with each class modeled by a mixture of Gaussian components (SL-EM). This method fits a mixture model onto each class data by expectation-maximization.

In addition to these supervised learning methods we implemented a number of benchmark semi-supervised learning algorithms. The first one is the semi-supervised EM algorithm introduced in [19] (SSL-EM). Briefly, this algorithm first, fits a Gaussian distribution onto each class data in the labeled data set, then, evaluates the posterior probabilities of the unlabeled samples for each class using the learned distributions, and finally, incorporates the unlabeled samples into the parameter-estimation process for each class by weighting them using their posterior probabilities. In this approach each unlabeled samples only contributes to the class for which the posterior is maximized. This process repeats until convergence and the resulting classifier is applied on the test data.

We also implemented two versions of the self-training method (SELF) with base learners ML and NB, respectively. Another algorithm we have included is the Co-training algorithm (CO-TR) implemented in two versions with base learners ML and NB respectively. For SELF and CO-TR we only include the better performing version in the experimental results.

One final approach we considered is an algorithm inspired by [8] (SSL-MOD). In this technique, similar to SSL-EM, we fit a Gaussian distribution for each class using labeled samples and use this model to classify unlabeled samples. The maximum of the class likelihood values are obtained for each sample. A two component Gaussian mixture model is fit onto the likelihood values in order to identify unlabeled samples with higher and lower likelihood values. We expect that unlabeled samples belonging to the observed classes will yield higher likelihood values whereas those from unobserved classes will yield low likelihood values. Then, we merge the unlabeled samples in the higher-likelihood group with the labeled samples to re-estimate the parameters of the classes. This process repeats until convergence and in the end another EM is performed on the samples remaning in the low-likelihood group to identify unobserved components. This technique is the only SSL technique, other than the proposed approach, that attempts to model unobserved classes. The proposed self-adjusting SSL approach is identifed by SA-SSL in this section.

B. Classifier Design and Evaluation

The labeled, unlabeled, and test data sets are generated as follows. We first divide the available labeled data into two and reserve one portion as test data. Then we further split the remaining portion into two as the labeled and unlabeled training data sets. During each split stratified sampling is used to make sure each class is proportionately represented in each subset. Some of the classes are considered unobserved and moved from the labeled set to the unlabeled set generating a non-exhaustive labeled data set. Both the unlabeled and test sets are exhaustive. The exact numbers for the number of unobserved classes and the proportions for the test, train and unlabeled sets are specified for each experiment below. We evaluate the performance of the classifiers using the overall classification accuracy and the average classification accuracies evaluated separately for observed and unobserved classes on the test set. We repeated this process ten times by generating ten random test/train/unlabeled splits and report the average accuracies along with the standard deviations.

The performance of the proposed SA-SSL algorithm is evaluated on three fronts: overall classification accuracy, classification accuracy for observed classes, classification accuracy for unobserved classes. To compute the classifier accuracy for unobserved classes each newly created component is assigned to the unobserved class having the majority of the samples in that component. Classification accuracy for each unobserved class is computed by the ratio of the total number of samples recovered by the corresponding components to the total number of samples in that class.

C. *Experiment 1: Pathogen Detection*

In this experiment a total of 2054 samples from 28 classes each representing a different bacteria serovar were considered. These are the type of serovars most commonly found in food samples. Each serovar is represented by between 40 to 100 samples where samples are the forward-scatter patterns characterizing the phenotype of a bacterial colony obtained by illuminating the colony surface by a laser light. Each scatter pattern is a gray level image characterized by a set of 22 features. More information about this dataset is available in [20]. We reserved 30% of the samples as test data, and used 30% of the remaining 70% as the labeled data set, while the rest remained as unlabeled. Four of the classes are considered unobserved and all of their samples are moved from the labeled set to the unlabeled set. So the non-exhaustive labeled set contains 24 classes and the exhaustive unlabeled and test data contains all of the 28 classes.

As the results in Table I suggest the proposed SA-SSL algorithm significantly outperforms all other techniques in terms of the overall classifier accuracy as well as classifier accuracies for observed and unobserved classes. In addition to classifying samples from unobserved components with a reasonable accuracy, the proposed approach also performs favorably compared to other techniques for classifying samples of observed components. On the average a total of 180 components and 150 unique covariance matrices were generated across twenty eight classes for this data set indicating that each class is modeled on the average with six components and about one sixth of the components shared covariance matrices with other components.

Method	Acc	Acc-O	Acc-U
SA-SSL	**0.81 (0.02)**	**0.80 (0.02)**	**0.84 (0.12)**
SSL-EM	0.64 (0.01)	0.75 (0.02)	0
SSL-MOD	0.67 (0.03)	0.74 (0.03)	0.26 (0.11)
SELF	0.59 (0.02)	0.70 (0.02)	0
CO-TR	0.60 (0.01)	0.72 (0.02)	0
SL-ML	0.62 (0.02)	0.73 (0.02)	0
SL-NB	0.52 (0.02)	0.62 (0.02)	0
SL-EM	0.30 (0.05)	0.35 (0.06)	0

Table I
AVERAGE OF 10 REPETITIONS, EACH RUN WITH DIFFERENT TEST/TRAIN/UNLABELED SPLITS OF THE BACTERIA DATASET. THE FIRST COLUMN SHOWS THE OVERALL ACCURACY ON THE TEST SAMPLES, SECOND AND THIRD COLUMNS SHOW THE ACCURACIES FOR THE OBSERVED AND UNOBSERVED CLASSES, RESPECTIVELY.

D. *Experiment 2: Multi-spectral Image Data Set*

We used the Flightline C1 multispectral image data set for this experiment. This is a 12-band multispectral image taken over Tippecanoe County, Indiana by the M7 scanner in June, 1966. There are eight classes, each class representing a different crop type. A total of 69,413 labeled pixels are available. More information about this multispectral imagery is available in [21]. In this data set we used 0.2% of all samples as the labeled data set, 5% as the unlabeled data set and the remaining samples are left for testing. One class is considered unobserved and moved from the labeled data set to the unlabeled data set, leaving a total of 121 labeled samples from seven classes for the non-exhaustive labeled set and around 3000 samples from all classes in the unlabeled set.

The proposed SA-SSL significantly outperforms all other techniques compared and recovers the one missing class with an almost perfect accuracy while achieving a fairly good accuracy for observed classes. A total of 20 components and 10 unique covariance matrices were generated across eight classes for this data set indicating that each class is modeled between two to three components and one half of the components shared covariance matrices with other components.

IV. CONCLUSIONS AND FUTURE WORK

We proposed a new framework for semi-supervised learning in partially-observed settings and presented results with two real-world datasets that favors the proposed approach in terms of improving the classifier performance even when there is a clear mismatch between the models that generated labeled and unlabeled data sets. In addition to more accurately classifying future samples of observed classes, this new approach can also discover unobserved classes and recover their samples with an acceptable accuracy.

Future research efforts will consider replacing the Gibbs sampler with scalable deterministic inference techniques, modifying the data model to accommodate text data,

Method	Acc	Acc-O	Acc-U
SA-SSL	**0.92 (0.01)**	**0.91 (0.01)**	**0.98 (0.01)**
SSL-EM	0.85 (0.01)	0.89 (0.01)	0
SSL-MOD	0.77 (0.06)	0.79 (0.07)	0.0
SELF	0.82 (0.01)	0.85 (0.01)	0
CO-TR	0.81 (0.02)	0.84 (0.02)	0
SL-ML	0.84 (0.01)	0.87 (0.01)	0
SL-NB	0.77 (0.02)	0.80 (0.02)	0
SL-EM	0.77 (0.01)	0.80 (0.01)	0

Table II
AVERAGE OF 10 REPETITIONS EACH RUN WITH DIFFERENT TEST/TRAIN/UNLABELED SPLITS OF THE MULTI-SPECTRAL IMAGE DATASET. THE FIRST COLUMN SHOWS THE OVERALL ACCURACY ON THE TEST SAMPLES, SECOND AND THIRD COLUMNS SHOW THE ACCURACIES FOR THE OBSERVED AND UNOBSERVED CLASSES, RESPECTIVELY.

and extending the developed framework to hierarchically-structured data sets to automatically associate newly discovered components and classes with higher level groups of classes.

ACKNOWLEDGMENT

The project was supported by Grant Number 5R21AI085531-02 from the National Institute of Allergy and Infectious Diseases (NIAID). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIAID or the National Institutes of Health.

REFERENCES

[1] T. Zhang and F. J. Oles, “A probability analysis on the value of unlabeled data for classification problems,” in *Proc. 17th International Conf. on Machine Learning*, 2000, pp. 1191–1198.

[2] F. Cozman and I. Cohen, “Unlabeled data can degrade classification performance of generative classifiers,” in *Fifteenth International Florida Artificial Intelligence Society Conference*, 2002, pp. 327–331. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.16.74

[3] Y. Guo, X. Niu, and H. Zhang, “An extensive empirical study on semi-supervised learning,” in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, dec. 2010, pp. 186 –195.

[4] G. McLachlan and K. Basford, *Mixture models. Inference and applications to clustering*. Marcel Dekker, 1988.

[5] V. Castelli and T. M. Cover, “The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter,” *Information Theory, IEEE Transactions on*, vol. 42, no. 6, pp. 2102–2117, Nov. 1996. [Online]. Available: http://dx.doi.org/10.1109/18.556600

[6] B. Shahshahani and D. Landgrebe, “The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 32, no. 5, pp. 1087–1095, 1994. [Online]. Available: http://dx.doi.org/10.1109/36.312897

[7] X. Zhu, “Semi-supervised learning literature survey,” Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005.

[8] D. J. Miller and J. Browning, “A mixture model and em-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1468–1483, 2003.

[9] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[10] M. D. Escobar and M. West, “Bayesian density estimation and inference using mixtures,” *Journal of the American Statistical Association*, vol. 90, pp. 577–588, 1994.

[11] T. S. Ferguson, “A Bayesian analysis of some nonparametric problems,” *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[12] D. J. Aldous, “Exchangeability and related topics,” in *École d’Été St Flour 1983*. Springer-Verlag, 1985, pp. 1–198, lecture Notes in Math. 1117.

[13] A. Dubey, I. Bhattacharya, M. K. Das, T. A. Faruquie, and C. Bhattacharyya, “Learning dirichlet processes from partially observed groups,” in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, 2011, pp. 141–150.

[14] M. Dundar, F. Akova, Y. Qi, and B. Rajwa, “Bayesian nonexhaustive learning for online discovery and modeling of emerging classes,” in *International Conference on Machine Learning (ICML’12) - June 26-July 1, 2012 - Edinburgh, Scotland*, 2012, to appear.

[15] Y. Teh, M. Jordan, M. Beal, and D. Blei, “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[16] H. Ishwaran and L. F. James, “Gibbs sampling methods for stick-breaking priors,” *Journal of the American Statistical Association*, vol. 96, no. 453, pp. pp. 161–173, 2001. [Online]. Available: http://www.jstor.org/stable/2670356

[17] F. Wood and M. J. Black, “A non-parametric Bayesian alternative to spike sorting,” *Journal of Neuroscience Methods*, vol. 173, pp. 1–12, 2008.

[18] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis, 3rd Edition*, 3rd ed. Wiley-Interscience, 2003.

[19] Q. Jackson and D. Landgrebe, “An adaptive classifier design for high-dimensional data analysis with a limited training data set,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 12, pp. 2664–2679, 2001.

[20] F. Akova, M. Dundar, V. J. Davisson, E. D. Hirleman, A. K. Bhunia, J. P. Robinson, and B. Rajwa, “A machine-learning approach to detecting unknown bacterial serovars,” *Statistical Analysis and Data Mining*, vol. 3, no. 5, pp. 289–301, 2010.

[21] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Newark, NJ: Wiley, 2005.