Dirichlet Process Based Active Learning and Discovery of Unknown Classes for Hyperspectral Image Classification

Hao Wu, *Student Member, IEEE*, and Saurabh Prasad, *Senior Member, IEEE*

***Abstract*—Active learning is an area of significant ongoing research interest for the classification of remotely sensed data, where obtaining efficient training data is both time consuming and expensive. The goal of active learning is to achieve high classification performance by querying as few samples as possible from a large unlabeled data pool. Traditional active learning frameworks all assume the existence of labeled samples for all classes of interest. However, in real-world applications, the unlabeled data pool may contain data from unknown classes that we are not aware of in advance, and a quick detection of them is useful for enriching our training set. In this scenario, traditional active learning methods may not effectively and rapidly detect the unknown classes. We proposed an active learning framework which provides robust classification performance with minimum manual labeling effort while simultaneously discovering unknown (missing) classes. The discovery of unknown classes is particularly suited to an active learning framework where an annotator is in the loop. A Dirichlet process mixture model is utilized in our proposed method to cluster the labeled and unlabeled samples as a whole. If unknown classes exist, they will emerge as new clusters which are different from other existing clusters occupied by known classes, and then, the proposed query strategy will give priority to querying samples in the new clusters. We present experimental results with hyperspectral data to show that our method provides better classification performance compared to existing active learning methods with or without unknown classes.**

***Index Terms*—Active learning, Dirichlet process mixture, hyperspectral image, unknown classes.**

I. Introduction

ACTIVE learning [1] is well motivated in many machine learning problems such as remote sensing [2]–[4], image retrieval [5], speech recognition [6], and natural language processing [7], [8], where unlabeled data are abundant but labeled data are very limited and annotation work is difficult, expensive, and time consuming. The main goal of an active learning algorithm is to achieve better classification performance by inducting as few samples as possible from an unlabeled data pool, which are labeled by an annotator and added to the training pool.

A key aspect of active learning is the construction of an effective query strategy which is designed to find the most informative samples and pose queries. A variety of query strategies have been created for active learning, including uncertainty sampling, query by committee (QBC) and expected model change, etc. [1], [2]. While there has been substantial work on active learning for classification, active learning with unknown class discovery has received considerably less attention. Traditional active learning systems assume that we have labeled data for every class of interest, even if the number of training samples initially available per class is small. However, we may encounter situations where we do not have labeled data for all classes, i.e., nothing is known about possibly new classes in the unlabeled data pool. This is common, particularly in remote sensing applications, where there may be unknown or new classes in unexplored geospatial areas. In such scenarios, traditional active learning will not work well with regard to the detection of the unknown classes. Thus, we aim to design an active learning system which is capable of detecting unknown classes as fast as possible and queries the most informative samples from them.

The Dirichlet process mixture model (DPMM) [9]–[12], a nonparametric Bayesian model, is well suited to data with complex mixture structures. A widely used DPMM is the infinite Gaussian mixture model (IGMM) [13] which overcomes the requirement of the number of mixture components in traditional Gaussian mixture modeling by assuming that data comes from a Gaussian mixture model (GMM) with an infinite number of mixture components. Due to the flexibility of DPMM, it has been used in many applications [14]–[17] for clustering and density estimation, etc. Thus, under DPMM, when a new class emerges, we will have one new cluster assigned to it or a few clusters due to the possible multimodal distribution of that class. We note that, with remote sensing images, due to spatial variability, new clusters do not have to be new classes—they may also come from a known class from a different spatial area than the labeled data. In [18] and [19], the new class detection problem has been addressed and solved by utilizing a DPMM, but detecting the new classes is just the first step to classification, so we carry it on to build an active learning framework with the ability of unknown class detection. We note that, although several classification paradigms exist for hyperspectral image analysis, including kernel-based machines [20], [21], sparse representation-based classifiers [22]–[25], etc., the focus of this paper is on Bayesian inference and probabilistic analysis techniques.

Manuscript received August 11, 2015; revised November 6, 2015, February 4, 2016, and February 21, 2016; accepted March 2, 2016. Date of publication April 28, 2016; date of current version June 1, 2016. This work was supported in part by the National Aeronautics and Space Administration under Grant NNX14AI47G. *(Corresponding author: Saurabh Prasad.)*

The authors are with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA (e-mail: saurabh.prasad@ieee.org).

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.2016.2552507

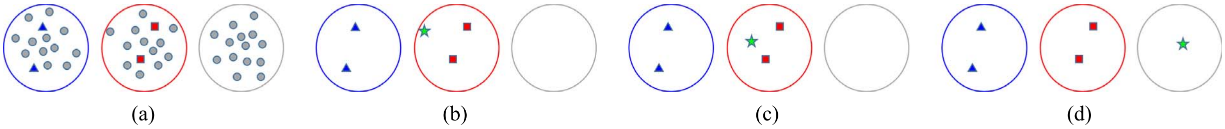


Fig. 1. Illustration comparing different methods for active learning with an unknown class. (a) Data distribution where polygons represent labeled instances from two known classes and circles represent unlabeled data from the known classes and one unknown class. (b) Sample queried by an uncertainty-based method [1], [7], marked as star. (c) Sample queried by a (global) ID-based method [7], marked as star. (d) Sample queried by the proposed LID-based method, marked as star.

In this paper, we integrate the new class detection problem into an active learning system for efficient classification. Although some work has been done on this problem based on DPMM in [26] which aims to discover rare classes in a pool of unlabeled images, the query strategy in [26] was efficient only for rare class discovery and assumed that every class (both known and unknown) occupies exactly one cluster. In our proposed framework, the unknown classes do not have to be rare. Additionally, our framework does not make an assumption about the number of clusters occupied by the known or unknown classes. This makes our method suitable for remote sensing, where classes may have multimodal distributions (for example, due to spatial variability within such images, classes can occupy more than one cluster).

The proposed query strategy—local information density (ID) (LID)—is built on a combination of a conventional uncertainty-based active query strategy and a local density generated by clustering, which makes it more suited for classification. Moreover, when new clusters are detected, we give priority to query data from new clusters which can contain data from either new classes or different spatial areas of known classes. That is, we wish to discover new classes and explore new clusters of existing classes as quickly as possible since both cases lead us to query data considered to be informative for the underlying active learning and image classification task. Fig. 1 gives an illustrative example demonstrating the effectiveness of our proposed approach in dealing with unknown classes, compared to traditional active learning approaches. Uncertainty sampling methods [1], [7] tend to query the most uncertain samples for the classifier based on the current training set. The queried samples are mainly located around the boundary between known classes, as illustrated in Fig. 1(b). In such a scenario, the unknown class will correspond to “certain samples” because it is easy to classify them as the “square” class based on the current training set. Thus, the uncertainty method cannot effectively detect the unknown class. The ID-based method, as proposed in [7], considers uncertainty and density at the same time, aiming to query uncertain samples that are representative of the unlabeled pool, i.e., have large global density. It helps to prevent the system from querying outliers which usually have high uncertainty but are not helpful to classification. A simulated result for ID is shown in Fig. 1(c), which failed to detect the unknown class because there is no guarantee that samples of unknown classes have high ID, which is, in fact, data dependent. For example, in a remote sensing scenario where the known classes are mainly vegetation classes and an unknown class is spectrally different, such as an urban class, the unknown class is more likely to have low ID because it is located far away from the known ones. Thus, the value of ID will not be large because the density term in the definition of ID, as described in Section II, has a relatively low value. Thus, in this case, ID-based methods cannot effectively detect the unknown class. However, the proposed LID method takes into account both uncertainty and a local density. It queries uncertain samples which have a high local density which is calculated within clusters generated by DPMM. When unknown classes are detected as new clusters by DPMM, some samples of unknown classes will have a high local density within the new clusters. Thus, by quickly identifying the new class as an emerging cluster, the LID-based method can direct active queries from that emerging cluster, as illustrated in Fig. 1(d).

Hyperspectral images provide the spectral information of objects over a wide range of the electromagnetic spectrum, usually with hundreds of bands, which yields precise characteristics of materials in the scene, compared to natural color images and multispectral images (often less than ten bands). However, the rich spectral information also comes with a big challenge, owing to the high dimensionality, particularly when using a statistical model for processing or analyzing such data. Dimensionality reduction is hence commonly undertaken for the feature extraction of hyperspectral images. Popular methods include unsupervised algorithms such as principal component analysis (PCA), Isomap [27], locally linear embedding (LLE) [28], [29], and locality preserving projection (LPP) [30] and supervised algorithms such as the Fisher linear discriminant analysis (LDA) (FLDA) and local Fisher discriminant analysis (LFDA) [31] and local tangent space alignment [29], [32]. Several of these methods are motivated by assumptions that the data reside on manifolds—a review of such methods for hyperspectral image analysis is provided in [33]. There are also semisupervised dimension reduction methods such as the semisupervised local Fisher discriminant analysis (SELF) [34], but they have not been developed for or studied in the context of unknown class discovery. In our framework, we employ SELF as a preprocessing, with the goal of maximizing separation between known classes in a lower dimensional subspace, while also trying to preserve the local structure of unlabeled samples which may contain data of unknown classes—i.e., we want to find a subspace where we can discriminate data from known classes without confusing data from unknown classes.

The remainder of this paper is organized as follows. Section II provides a brief literature review about related work. The proposed framework is described in detail in Section III. A description of the data sets used for validation is given in Section IV. The experimental setup and results validating the proposed approach are detailed in Section V. Section VI summarizes the key ideas and experimental results in this paper and provides concluding remarks.

II. Related Work

A. Active Learning

In general, an active learning system starts with a small labeled data set $\mathcal{L}$. It iteratively selects the most informative samples from the unlabeled data set $\mathcal{U}$, queries their labels (which is done by a human annotator in real-world application), and adds them to $\mathcal{L}$, aiming to improve the classification performance with the least number of queried samples. Three main scenarios for the active learning problem have been considered in the literature: 1) pool-based sampling; 2) stream-based selective sampling; and 3) membership query synthesis. Since the pool-based scenario is the most widely used in remote sensing, we only consider this scenario where samples are queried from a large unlabeled data pool and labeled by a human annotator. The key task in active learning involves evaluating the informativeness of each unlabeled sample based on an appropriate query strategy $\phi(\cdot)$.

For posterior probability-based active learning methods, one of the most common query strategies for evaluating informativeness is uncertainty sampling [35] which includes least confidence (LC), breaking ties (BT), entropy, etc. The LC strategy queries the instance for which the current model has the LC in its most likely labeling, which is defined as

$$\phi_{\mathrm{LC}}(\boldsymbol{x}) = 1 - P(\hat{y}|\boldsymbol{x}) \tag{1}$$

where $\hat{y}$ is the most probable class label for $\boldsymbol{x}$, i.e., $\hat{y} = \arg\max_y P(y|\boldsymbol{x})$.

BT queries the instance with the smallest difference between posteriors for its two most likely labelings, which is defined as

$$\phi_{\mathrm{BT}}(\boldsymbol{x}) = P(\hat{y}_1|\boldsymbol{x}) - P(\hat{y}_2|\boldsymbol{x}) \tag{2}$$

where $\hat{y}_1$ and $\hat{y}_2$ are the first and second most probable class labels for $\boldsymbol{x}$ under the current model.

Entropy-based uncertainty sampling queries the instance which has the largest entropy

$$\phi_E(\boldsymbol{x}) = -\sum_{j=1}^{N_c} P(y_j|\boldsymbol{x}) \log P(y_j|\boldsymbol{x}) \tag{3}$$

where N_c is the total number of possible class labels.

QBC [36], another active learning framework, measures informativeness based on the degree of disagreement between a committee of classification models. Another framework is the expected model change, which queries the instance that will result in the biggest change to the current model if added to the labeled set. An example of this framework is the expected gradient length (EGL) [37].

However, the conventional active learning frameworks such as those described earlier have a drawback that they are prone to querying outliers which typically have high uncertainty and large disagreement between a committee. For uncertainty sampling, outliers and the most uncertain samples lie on the classification boundary, and they are not "representative" of other samples in the distribution; thus, knowing their labels is unlikely to improve the classification performance on the data as a whole. Representativeness measures the ability of a sample to express the distribution of the whole data. Thus, outliers with high informativeness are not representative of the unlabeled samples—as a result, they are not beneficial for classification. Similarly, QBC and EGL will spend time querying possible outliers simply because they are controversial or they are expected to impart the most significant change to the model. This is commonly seen during the initial query steps of active learning, when we do not have enough labeled samples, particularly for classifiers built from generative models.

To address this, ID as a query strategy was proposed in [7], which aims to combine informativeness and representativeness, and is defined as

$$\phi_{\mathrm{ID}}(\boldsymbol{x}) = \phi_E(\boldsymbol{x}) \left(\frac{1}{|\mathcal{U}|-1} \sum_{\boldsymbol{x}^{(u)} \in \mathcal{U} \backslash \boldsymbol{x}} \mathrm{sim}\left(\boldsymbol{x}, \boldsymbol{x}^{(u)}\right) \right)^{\beta} \tag{4}$$

which implies that the informativeness of an unlabeled sample $\boldsymbol{x}$ is weighted by its average similarity to all the other samples in $\mathcal{U}$ with a parameter β that controls the relative importance of the density term. $|\mathcal{U}|$ in (4) denotes the number of samples in the candidate data set $\mathcal{U}$. ϕ_E in (4) serves as the "base" informativeness measure which can be an uncertainty criterion or QBC, etc. Here, we choose entropy to be the "base" measure. Two commonly used similarity measures are the exponential Euclidean distance

$$\mathrm{sim}\left(\boldsymbol{x}, \boldsymbol{x}^{(u)}\right) = \exp\left(-\frac{\left\|\boldsymbol{x} - \boldsymbol{x}^{(u)}\right\|}{\sigma^2} \right)$$

and the cosine similarity

$$\mathrm{sim}\left(\boldsymbol{x}, \boldsymbol{x}^{(u)}\right) = \frac{\boldsymbol{x} \cdot \boldsymbol{x}^{(u)}}{\|\boldsymbol{x}\| \times \left\|\boldsymbol{x}^{(u)}\right\|}.$$

In our experiments, we find that the exponential Euclidean distance performs slightly better than the cosine distance. Thus, the exponential Euclidean distance is employed in this paper. However, ID has a drawback—samples with high ID may not be sufficiently representative for every class, even though they are representative for the entire unlabeled data set. Thus, such a query scheme may focus on querying data from only a few classes which have much higher density than the other classes, making it less competitive for classification and new class discovery.

B. Semisupervised Dimension Reduction

In active learning problems, we typically start with a small number of labeled samples (i.e., training data) and a large amount of unlabeled samples from which we will choose the most informative ones to be labeled by a human annotator. For high-dimensional hyperspectral data analysis, particularly when we do not have enough labeled samples, a feature reduction preprocessing is often utilized to reduce the number of features. Due to the lack of sufficient labeled samples, supervised dimension reduction methods such as FLDA [38] and its localized version—LFDA [31]—do not work well since they tend to suffer from overfitting. In addition, potentially unknown classes may emerge from the unlabeled data pool, in which case the projection computed by supervised dimensionality reduction methods based on labeled data may not preserve the local structure of the unknown classes. Although unsupervised dimension reduction methods such as PCA, LLE [28], and LPP [30] can be used to avoid this issue, they discard the discriminative information carried by the labels. Thus, a natural approach to take both the labeled and unlabeled data into account is to use a semisupervised dimension reduction algorithm. In [34], SELF was proposed by combining the information in both the labeled data via LFDA and the unlabeled data via PCA or LPP. Since LPP preserves locality in an unsupervised manner and it is less likely to mix unknown classes with the known classes, we choose LPP on the unlabeled data when implementing SELF. Thus, SELF separates the labeled samples in different classes based on their labels and tries to preserve the local structure of unlabeled samples at the same time.

Let $\boldsymbol{x}_i \in \mathbb{R}^d$ be the ith sample vector (e.g., pixel) and $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \in \mathbb{R}^{d\times n}$ be the data matrix. Let $\boldsymbol{y} \in \mathbb{R}^r$ $(1 \le r \le d)$ be the low-dimensional representation of $\boldsymbol{x}$ via a projection matrix $\boldsymbol{T} \in \mathbb{R}^{d\times r}$: $\boldsymbol{y} = \boldsymbol{T}^\top \boldsymbol{x}$. In general, many linear dimension reduction algorithms can be formulated as the following optimization problem [34]:

$$\boldsymbol{T}^{(\mathrm{OPT})} = \mathop{\arg\max}_{\boldsymbol{T}\in\mathbb{R}^{d\times r}} \left[\mathrm{tr}\left(\boldsymbol{T}^\intercal \boldsymbol{B}\boldsymbol{T}(\boldsymbol{T}^\top \boldsymbol{C}\boldsymbol{T})^{-1}\right)\right] \tag{5}$$

where $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ encodes the quantity that we want to increase (e.g., between-class separability) and $\boldsymbol{C} \in \mathbb{R}^{d\times d}$ corresponds to the quantity that we want to decrease (e.g., within-class scatter).

In both LPP and LFDA, an affinity matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is used to quantify the affinity between samples, with A_{ij} defined as

$$A_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma_i\sigma_j}\right) \tag{6}$$

where σ_i represents the local scaling around $\boldsymbol{x}_i$ defined by $\sigma_i = \|\boldsymbol{x}_i - \boldsymbol{x}_i{}^{(k)}\|$ where $\boldsymbol{x}_i^{(k)}$ is the kth nearest neighbor of $\boldsymbol{x}_i$ in the original feature space.

In LPP, $\boldsymbol{B}$ and $\boldsymbol{C}$ are defined as the normalization matrix and the local scatter matrix

$$\boldsymbol{S}^{(n)} = \boldsymbol{X}\boldsymbol{D}^{(n)}\boldsymbol{X}^\top \tag{7}$$

$$\boldsymbol{S}^{(l)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}^{(l)}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \tag{8}$$

where $\boldsymbol{D}^{(n)}$ is the $n \times n$ diagonal matrix defined with $D_{ii}^{(n)} = (1/n)\sum_{j=1}^{n} A_{ij}$ and $W_{ij}^{(l)} = (1/n)A_{ij}$.

In LFDA, $\boldsymbol{B}$ and $\boldsymbol{C}$ are the local between-class scatter matrix and the local within-class scatter matrix defined as

$$\boldsymbol{S}^{(lb)} = \sum_{i,j=1}^{n'} \frac{W_{ij}^{(lb)}}{2}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \tag{9}$$

$$\boldsymbol{S}^{(lw)} = \sum_{i,j=1}^{n'} \frac{W_{ij}^{(lw)}}{2}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \tag{10}$$

where n' is the number of labeled samples and $\boldsymbol{W}^{(lb)}, \boldsymbol{W}^{(lw)}$ are $n' \times n'$ matrices with

$$W_{ij}^{(lb)} = \begin{cases} A_{ij}\left(1/n' - 1/n'_{y_i}\right), & \text{if } y_i = y_j \\ \frac{1}{n'}, & \text{if } y_i \neq y_j \end{cases} \tag{11}$$

$$W_{ij}^{(lw)} = \begin{cases} \frac{A_{ij}}{n'_{y_i}}, & \text{if } y_i = y_j \\ 0, & \text{if } y_i \neq y_j \end{cases} \tag{12}$$

where n'_{y_i} denotes the number of samples in class y_i.

In SELF, the labeled data contribute to the projection via the LFDA term, while an LPP-like term is used to preserve the locality of the unlabeled data. B and $\boldsymbol{C}$ are the regularized local between-class scatter matrix and the regularized local within-class scatter matrix defined as

$$\boldsymbol{S}^{(rlb)} = (1-\gamma)\boldsymbol{S}^{(lb)} + \gamma\boldsymbol{S}^{(n)} \tag{13}$$

$$\boldsymbol{S}^{(rlw)} = (1-\gamma)\boldsymbol{S}^{(lw)} + \gamma\boldsymbol{S}^{(l)} \tag{14}$$

where $\gamma \in [0, 1]$ is a tradeoff parameter which controls the importance of LFDA and LPP in SELF.

C. Active Learning With New Class Discovery

New class (novelty) detection can be described as the identification of new or "unknown" data that a machine learning system was not aware of during training. The ability to detect new classes can have a significant impact in remote sensing applications, where the unlabeled data may contain information about objects that were not present in the labeled data.

Since DPMMs are capable of fitting data with an unknown number of mixtures, it is possible to differentiate between known and unknown classes by learning the clustering structure of the labeled and unlabeled data. If unlabeled data contain new classes, they will be assigned to new clusters that are different from those of the known classes. This has been done in [18] and [19]. Since our final goal is classification, we need to acquire labeled data of new classes after detecting them. Thus, in this paper, we embed the new class discovery problem in an active learning framework, aiming to improve classification performance with the least possible query effort for both known and unknown classes.

Recent work [26] for this problem based on DPMM is aimed at discovering rare classes in a pool of unlabeled images. The query strategy in [26] was constructed on the clustering result, which makes it efficient in class discovery. However, it is assumed that the unknown classes are rare and every class (including known and unknown) occupies one cluster, which are not valid assumptions in many remote sensing applications. First, although rare unknown classes can be commonly encountered in various applications, with remotely sensed image analysis over wide geographic areas, we can expect scenarios where unknown classes can also be prevalent (not rare). Second, for remotely sensed hyperspectral images, the properties of some classes may have large spatial variability, which makes their distribution possess a complex form (e.g., a multimodal distribution). Thus, the initial training set of a known class

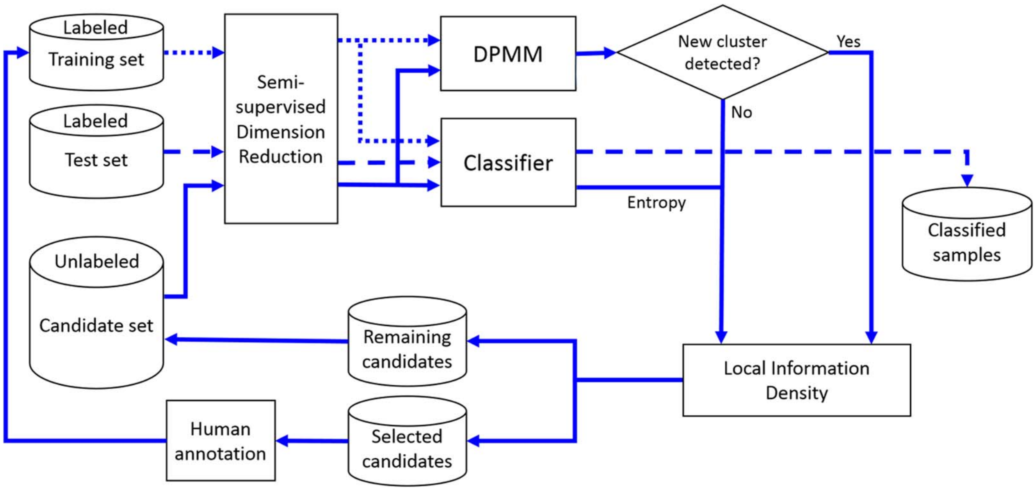


Fig. 2. Flowchart of the proposed active learning framework.

may only contain samples from a specific region with low spatial variability, while the unlabeled set may contain candidate samples from other unexplored regions with a different distribution than the samples in the training set. In this case, the samples in the unlabeled pool can form a new cluster by DPMM which is different from the cluster occupied by the samples of the same class in the training set. It is expected that these unlabeled samples are also useful for classification. The method that we propose not only detects new classes but also detects unexplored areas of known classes which contain useful information for classification not currently available in the training data set. It is important to note that not all new classes may be quickly and effectively detected. Note that there may be scenarios where some new classes may not be quickly detected (for example, if their spectra are very similar to some known class, they may merge into the same cluster).

III. Proposed Framework

The flowchart of the proposed active learning system with a query strategy based on LID is shown in Fig. 2. In the proposed framework, SELF is used for the semisupervised dimensionality reduction of the hyperspectral imagery. We expect SELF to be particularly appropriate for our problem since the unsupervised (LPP) component in SELF will ensure that dominant information pertinent to missing classes is not lost in an otherwise supervised dimensionality reduction approach (with the intuition that not accounting for missing classes in a completely supervised feature reduction approach will result in suboptimal transformations wherein information about missing classes may be potentially lost). Following this, the unlabeled data are clustered via DPMM, and local density for each candidate is calculated based on the clustering result. At the same time, entropy for each candidate is computed from the posterior provided by the classifier. Queries are then made based on LID which is computed from local density and entropy. Finally, the queried samples are added to the training set and removed from the candidate set.

In the following discussion, clustering is applied to data $\boldsymbol{X}$ in a lower dimensional subspace obtained from SELF. Compared to a static projection (such as LDA) as is commonly utilized in traditional classification, we have a dynamic projection which is refined after each step of active learning. As more data from both known and unknown classes are labeled and added to the training set, we will reach a better subspace given by SELF.

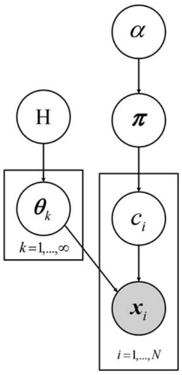


Fig. 3. Graphical model representation of IGMM, where the shaded and unshaded nodes indicate observed and latent variables, respectively.

A. Clustering Based on DPMM

GMMs can successfully capture the complex multimodal statistical structure of data, including remotely sensed data [39], audio data [40], etc. However, GMM assumes the number of Gaussian components to be known and uses the expectation–maximization algorithm for parameter inference. IGMM does not have this limitation by assuming the number of Gaussian components to be infinity. Given the data, the number of components can be inferred automatically by Markov chain Monte Carlo (MCMC) [41] sampling methods. IGMM, as defined in [13], is a special case of a Dirichlet process mixture, where the mixture components are assumed to be Gaussian distributed. In our framework, DPMM is used to cluster data in both the labeled set $\mathcal{L}$ and unlabeled set $\mathcal{U}$, as shown in Fig. 2. The graphical representation for the generative model of IGMM is depicted in Fig. 3 where the observations $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ are generated from a GMM with an infinite number of Gaussian components. In our active learning framework, $\boldsymbol{x}_i$ corresponds to a pixel in the labeled set $\mathcal{L}$ and unlabeled set $\mathcal{U}$, and c_i is the cluster assignment for $\boldsymbol{x}_i$ [14], [15].

The Dirichlet process is a distribution over countably infinite random measures on a parameter space $\boldsymbol{\Theta}$. We write $G \sim \mathrm{DP}(\alpha, H)$ to indicate

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k} \tag{15}$$

where $\boldsymbol{\theta}_k$ denotes independent random variables distributed according to H and $\delta_{\boldsymbol{\theta}_k}$ is an indicator function centered at $\boldsymbol{\theta}_k$ (zero elsewhere except for $\delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k) = 1$). The weights $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_\infty)$ are generated via a stick-breaking construction process [42] denoted by $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$ (GEM stands for Griffiths, Engen, and McCloskey [43]) which is defined as

$$\beta_{\mathrm{k}} \sim \mathrm{Beta}(1, \alpha);\ \pi_{\mathrm{k}} = \beta_{\mathrm{k}} \prod_{\mathrm{l}=1}^{\mathrm{k}-1} (1 - \beta_l), \quad k = 1, 2, \ldots, \infty \tag{16}$$

where $\pi_k \in [0, 1]$ and $\sum_{k=1}^{\infty} \pi_k = 1$. The concentration parameter α in the Dirichlet process encodes how concentrated the samples from a DP will be around the base measure H [12].

Given $\boldsymbol{\pi}$, an indicator variable c_i, denoting the Gaussian component that generates $\boldsymbol{x}_i$, is obtained through

$$c_i | \boldsymbol{\pi} \sim \mathrm{Multinomial}(\cdot | \boldsymbol{\pi}). \tag{17}$$

Finally, $\boldsymbol{x}_i$ is generated from the Gaussian distribution indicated by c_i

$$\boldsymbol{x_i} | c_i = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),\ k = 1, 2, \ldots, \infty. \tag{18}$$

For simplicity, we write $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ to denote parameters of the kth Gaussian component in the following discussion.

We put a conjugate prior on $\boldsymbol{\theta}_k$ as suggested by $\boldsymbol{\theta}_k \sim H$ in (15), which is chosen to be the normal-inverse-Wishart (NIW) distribution

$$p(\boldsymbol{\theta}_k | H) = p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k | H) := \mathrm{NIW}(\boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Lambda}_0, \nu_0) \tag{19}$$

which is equivalent to

$$\begin{aligned} \boldsymbol{\mu}_k &\sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{\boldsymbol{\Sigma}_k}{\kappa_0}\right) \\ \boldsymbol{\Sigma}_k &\sim \mathcal{IW}\left(\boldsymbol{\Lambda}_0^{-1}, \nu_0\right) \end{aligned}$$

where $\boldsymbol{\mu_0}$ is the mean vector and κ_0 is the relative precision for the Gaussian prior on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_0^{-1}$ is the scale matrix and ν_0 is the degrees of freedom for the inverse Wishart prior on $\boldsymbol{\Sigma}_k$. The inverse Wishart distribution is defined as

$$p\left(\boldsymbol{\Sigma}_k | \boldsymbol{\Lambda}_0^{-1}, \nu_0\right) = \frac{|\boldsymbol{\Lambda}_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 p}{2}} \Gamma_d\left(\frac{\nu_0}{2}\right)} |\boldsymbol{\Sigma}_k|^{-\frac{\nu_0 + d + 1}{2}} e^{-\frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}_k^{-1}\right)} \tag{20}$$

where $\Gamma_d(\cdot)$ is the d-dimensional gamma function. This NIW distribution is a conjugate prior for $\boldsymbol{\theta}_k$, which means that the posterior will have the same form given that the likelihood is Gaussian.

The generative model for IGMM described earlier generates observations given model parameters. Thus, data can be clustered according to their indicator variables $\boldsymbol{C} = \{c_1, \ldots, c_N\}$. During inference, given data $\boldsymbol{X}$, our goal is to infer the model parameters, i.e., the cluster assignments $\boldsymbol{C} = \{c_1, \ldots, c_N\}$ and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, by obtaining samples from their posterior distribution using MCMC sampling approaches [41]. Since the direct sampling of all the parameters is difficult, thus, in our implementation, we use a MCMC algorithm known as Gibbs sampling [41], which iteratively obtains a new value for each parameter from its posterior distribution conditioned on the current values of all the other parameters. Like other MCMC sampling methods, the Gibbs sampler can generate a sequence of samples for each parameter, and upon convergence, these samples will approximate the posterior distribution.

Specifically, in each iteration of Gibbs sampling, $\boldsymbol{\theta}_k$ ($k = 1, \ldots, K$, where K is the current number of clusters) is updated by sampling from its posterior distribution

$$p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k | \boldsymbol{X}, \boldsymbol{C}) = \mathrm{NIW}\left(\boldsymbol{\mu}_{n_k}, \kappa_{n_k}, \boldsymbol{\Lambda}_{n_k}, \nu_{n_k}\right) \tag{21}$$

with

$$\begin{aligned} \boldsymbol{\mu}_{n_k} &= \frac{\kappa_0}{\kappa_0 + n_k} \boldsymbol{\mu}_0 + \frac{n_k}{\kappa_0 + n_k} \bar{\boldsymbol{x}}, \\ \kappa_{n_k} &= \kappa_0 + n_k \\ \nu_{n_k} &= \nu_0 + n_k \\ \boldsymbol{\Lambda}_{n_k} &= \boldsymbol{\Lambda}_0 + \boldsymbol{S} + \frac{\kappa_0 n_k}{\kappa_0 + n_k} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^\top \end{aligned}$$

where

$$\bar{\boldsymbol{x}} = \frac{1}{n_k} \sum_{c_i = k} \boldsymbol{x}_i$$

$$\boldsymbol{S} = \sum_{i=1}^{n_k} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$$

and n_k is the number of data instances assigned with the kth cluster.

Similarly, c_i ($i = 1, \ldots, N$) is updated by sampling from its posterior distribution

$$p(c_i = k | \boldsymbol{C}_{-i}, \boldsymbol{X}, \boldsymbol{\Theta}, \alpha) \propto p(\boldsymbol{x}_i | c_i = k, \boldsymbol{\Theta}) p(c_i = k | \boldsymbol{C}_{-i}, \alpha) \tag{22}$$

where the prior term

$$p(c_i = k | \boldsymbol{C}_{-i}, \alpha) = \begin{cases} \frac{n_{k,-i}}{N - 1 + \alpha}, & \text{if } k \in \boldsymbol{C}_{-i} \\ \frac{\alpha}{N - 1 + \alpha}, & \text{otherwise} \end{cases} \tag{23}$$

with $C_{-i} = \{c_1, \ldots, c_{k-1}, c_{k+1}, \ldots, c_N\}$ referring to all the indicator variables, except c_i, and $n_{k,-i}$ referring to the number of data instances assigned to the kth cluster, excluding $\boldsymbol{x}_i$. Moreover, the likelihood term

$$p(\boldsymbol{x}_i | c_i = k, \boldsymbol{\Theta}) = \begin{cases} p(\boldsymbol{x}_i | \boldsymbol{\theta}_k), & \text{if } k \in \boldsymbol{C}_{-i} \\ \int p(\boldsymbol{x}_i | \boldsymbol{\theta}) p(\boldsymbol{\theta} | H) d\boldsymbol{\theta}, & \text{otherwise.} \end{cases} \tag{24}$$

After simplification, the integration in (24) turns out to be a d-dimensional Student's t-distribution

$$t_{\nu_0 - d + 1}\left(\boldsymbol{\mu}_0, \frac{\boldsymbol{\Lambda}_0(\kappa_0 + 1)}{\kappa_0(\nu_0 - d + 1)}\right). \tag{25}$$

A standard d-dimensional Student's t-distribution with parameters ν (degrees of freedom), $\boldsymbol{\mu}$ (location vector), and $\boldsymbol{\Sigma}$ (scale matrix) is defined as

$$t_v(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\Gamma\left[\frac{(\nu+d)}{2}\right]}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{d}{2}}\pi^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\left[1+\frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{\frac{(\nu+d)}{2}}}. \tag{26}$$

Finally, we assign cluster labels for each unlabeled instance in $\mathcal{U}$ using the sample from the last iteration of the Gibbs sampling process. It is possible for a few outliers to appear from very small-sized clusters. Hence, clusters with too few samples (empirically set to ten samples) will be discarded at the current query step in our implementation. We should note that these samples are temporarily discarded and may appear to reside in large clusters again in later query steps.

In [19], DPMM and another nonparametric Bayesian clustering method, the infinite warped mixture model (IWMM) which can handle heavy-tailed distributions, were successfully used for new class detection. IWMM [44] assumes an implicit mapping from some latent space, where the latent representation (which can be inferred by MCMC) of data is ideally Gaussian distributed, to the observed space where data (e.g., pixels of images) may have heavy-tailed distribution. Thus, we can get a better clustering result in the latent space if the assumption holds, which is verified in [44]. However, the latent representations of the test samples cannot be derived due to the lack of an explicit mapping function, which makes IWMM not an appropriate method in our framework.

B. Query Strategy Based on LID

The ID described in Section II has a drawback that samples with high ID may not be representative of every class even though they are representative of the entire unlabeled data set. Hence, it may focus on querying data from only a few classes which have much higher density than the other classes, making it less competitive. In this paper, a LID is proposed here to address this problem, which aims to query samples that are both informative and representative of every class in $\mathcal{U}$. In the ID strategy, the density of $\boldsymbol{x}$ is calculated by its average similarity to all the other samples in the unlabeled data pool $\mathcal{U}$. In the proposed LID strategy, the density of $\boldsymbol{x}$ is calculated by its average similarity to its neighboring samples, and the neighborhood is defined by clusters generated by DPMM. Since both labeled samples in $\mathcal{L}$ and unlabeled samples in $\mathcal{U}$ are clustered together, some clusters generated by DPMM contain both labeled and unlabeled samples, and the others only contain unlabeled samples. Those containing only unlabeled samples are considered to be new clusters. These samples in new clusters are either from unknown classes or known classes in unexplored regions of an image due to the spatial variability of some classes.

Given the cluster assignments of all the unlabeled candidates in $\mathcal{U}$, the local density for an unlabeled candidate $\boldsymbol{x}$ with cluster label c can be computed as

$$\phi_{\mathrm{LD}}(\boldsymbol{x}) = \frac{1}{|\mathcal{U}_c|-1}\sum_{\boldsymbol{x}^{(u)}\in\mathcal{U}_c\backslash\boldsymbol{x}} \mathrm{sim}\left(\boldsymbol{x},\boldsymbol{x}^{(u)}\right) \tag{27}$$

where $\mathcal{U}_c$ represents all the unlabeled samples with the same cluster assignment c and $|\mathcal{U}_c|$ denotes the size of cluster c.

When new clusters emerge, in many applications, it is highly desirable to prioritize new clusters, to enable fast discovery. Hence, local densities are computed only for the samples in new clusters, while the local densities for all the other samples are set to be zero. If no new cluster is found, local densities will be calculated for every unlabeled sample in $\mathcal{U}$. Thus, our LID is formulated as

$$\phi_{\mathrm{LID}}(\boldsymbol{x}) = \begin{cases}\phi_{\mathrm{LD}}(\boldsymbol{x})\mathrm{I}_{\mathrm{new}}, & \text{if new cluster exists}\\ \phi_E(\boldsymbol{x})\left(\phi_{\mathrm{LD}}(\boldsymbol{x})\right)^{\beta}, & \text{otherwise}\end{cases} \tag{28}$$

where the indicator function $\mathrm{I}_{\mathrm{new}}(\boldsymbol{x})$ equals to 1 only when $\boldsymbol{x}$ comes from a new cluster and 0 otherwise. As in ID, we choose entropy $\phi_E(\cdot)$ to be the base measure. The parameter β which controls the relative importance of the local density term can simply be set to be 1 to make the two terms equally important. In our implementation, the value of β is set in an adaptive way such that the relative importance of the local density term $\phi_{\mathrm{LD}}(\boldsymbol{x})$ gets lower as more and more samples are labeled. Thus, we use a simple scheme to achieve this as follows:

$$\beta = \frac{|\mathcal{U}|}{|\mathcal{U}|+|\mathcal{L}|}. \tag{29}$$

The ID-based query strategy defined in (4) combines the entropy with a global density term which is calculated by averaging the similarities between $\boldsymbol{x}$ and all the other samples in the unlabeled pool $\mathcal{U}$. It aims to query globally representative samples without emphasizing new classes. For our proposed LID-based strategy queries, when new clusters are detected, it queries samples from the new clusters. When no new clusters exists, it queries uncertain samples with large local density computed in local clusters, which are locally representative and more beneficial for classification.

With the setup described earlier, the outline of the proposed active learning framework is summarized in Algorithm 1.

Algorithm 1 *Active learning with unknown class discovery*

```
Inputs:
    — Initial labeled set $\mathcal{L}$
    — Unlabeled candidate pool $\mathcal{U}$
    — Batch size $B$

while stopping criterion has not been met do
    Dimension reduction via SELF for data in $\mathcal{L}$ and $\mathcal{U}$.
    Cluster data in $\mathcal{L} \cup \mathcal{U}$ via DPMM:
    for each Gibbs sampling iteration do
        Update model parameters $\boldsymbol{\Theta}$ via (21).
        Update cluster assignments $\boldsymbol{C}$ via (22).
    end for
```

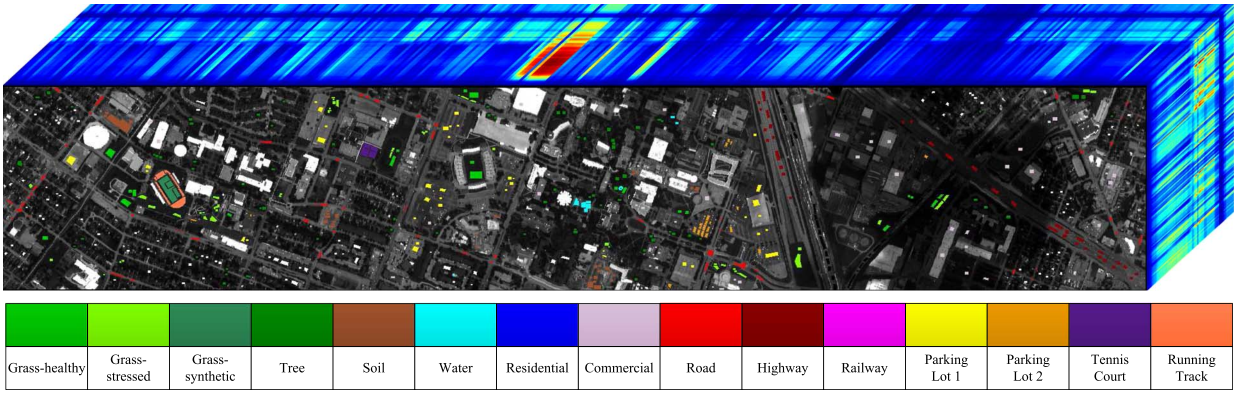


Fig. 4. Color image of UH data set with ground truth for 15 classes in different colors.

```
Train the classifier using the current training set $\mathcal{L}$ and apply it to $\mathcal{U}$.
for each candidate $\boldsymbol{x}_i \in \mathcal{U}$ do
    Compute $P(y|\boldsymbol{x}_i)$ given by the classifier.
    Evaluate entropy $\phi_E(\boldsymbol{x}_i)$ via (3).
    Calculate $\phi_{\mathrm{LD}}(\boldsymbol{x}_i)$ via (27) based on the clustering results.
    Calculate $\phi_{\mathrm{LID}}(\boldsymbol{x}_i)$ via (28).
end for
Rank the candidates based on their LIDs.
The top $B$ candidates are selected to be labeled by an annotator, then removed from $\mathcal{U}$ and added to $\mathcal{L}$.
end while
```

Outputs:
— Classifier trained on the updated labeled set $\mathcal{L}$

IV. Experimental Data

A. Data Sets

The first data set used in this paper, the University of Houston (UH) hyperspectral image, was acquired by the NSF-funded National Center for Airborne Laser Mapping over the UH campus and the neighboring urban area using the ITRES Compact Airborne Spectrographic Imager 1500 hyperspectral imager. The hyperspectral image has 15 classes and contains 144 spectral bands over the 364–1046-nm wavelength range. It has a spatial dimension of 1905 × 349 with a spatial resolution of 2.5 m. Fig. 4 shows the true color image of the UH data set with ground truth for all the 15 classes, and Fig. 5 shows the corresponding mean spectral signatures (radiance) for each class. This data set contains a training set with 2832 labeled pixels and a validation set with 12 197 labeled pixels. In the following active learning experimental setup, the initial labeled set $\mathcal{L}$ and candidate set $\mathcal{U}$ are chosen from the training set, and the test set is chosen from the validation set.

The second data set, the Indian Pines hyperspectral image, was acquired using the ProSpecTIR instrument in May 2010 over an agriculture area in Indiana, USA. The image has a 1342 × 1287 spatial dimension with 2-m spatial resolution. It consists of 360 spectral bands over the 400–2500-nm wavelength range. The 19 classes contain agriculture fields with different residue covers. Fig. 6 shows the true color image of the data set with the corresponding ground truth, and Fig. 7 shows the mean spectral signatures (reflectance) of these 19 classes.

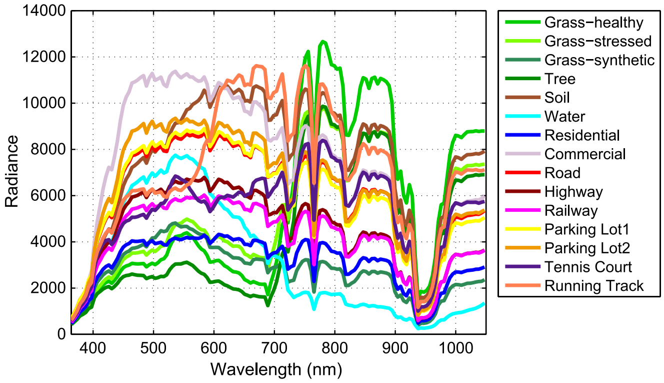
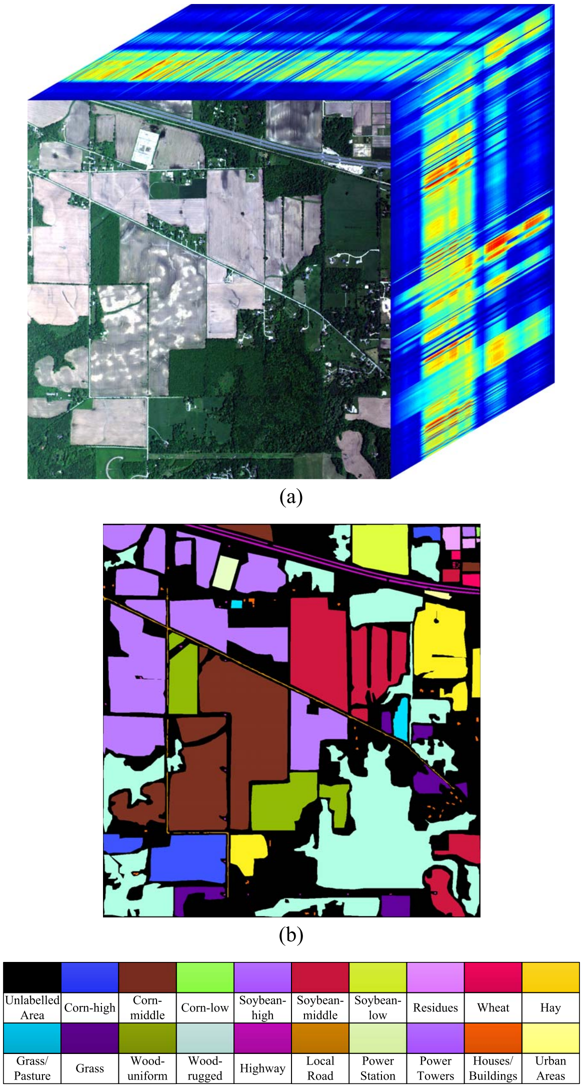


Fig. 5. Mean spectral signatures of 15 classes from the UH data set where different colors correspond to different classes in Fig. 4.

V. Experimental Setup and Results

For dimension reduction, we use SELF to reduce the dimensionality of the two data sets to 15, and a constant $\gamma = 0.5$ in SELF is used to balance the effects of LFDA and LPP. To choose the number of features to retain after SELF, we use a cross-validation method (leave one out) on the training data, and we find 15 to be an optimal value for our data sets. For the active learning setup, 20 samples per class are randomly selected from the data set to be the initial training set, and 150 samples per class are randomly selected to be the candidate data set. The initial number of clusters in DPMM is set to be the number of known classes, and we take the sample mean and sample covariance of the training samples as the parameter initialization of each Gaussian component, which is a reasonable choice for 20 samples per cluster when the dimensionality of the data is 15. We run active learning for 80 iterations with a batch size $B = 5$ (number of samples queried at each iteration). For validation, 200 samples per class are chosen to be the test



(a)

(b)

(c)

Fig. 6. Indian Pines data set: (a) True color image and (b) ground truth with class names in (c).

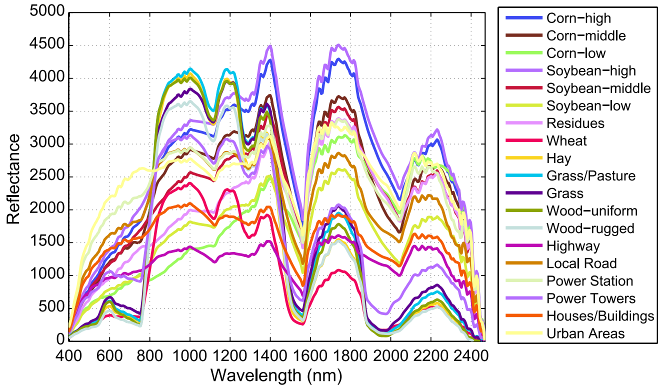


Fig. 7. Mean spectral signatures of the 19 classes in the Indian Pines data set.

data set on which the classifier is evaluated for each query step of active learning. For each query step of active learning, labeled and unlabeled samples are clustered via DPMM using 100 iterations of Gibbs sampling per run.

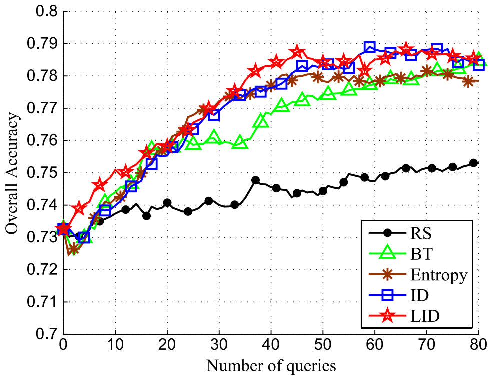


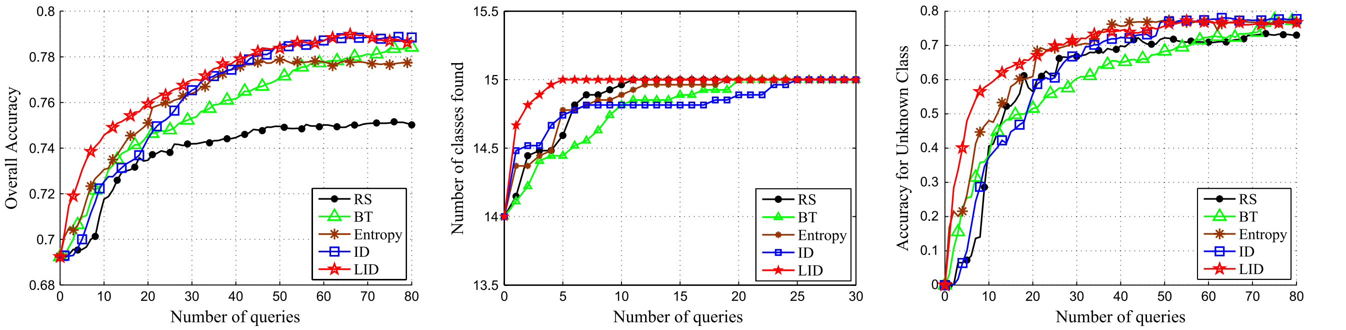
Fig. 8. Learning curves of overall accuracy for all query strategies without unknown classes for the UH data set.

To simulate scenarios where unknown classes occur in the unlabeled candidate pool, randomly selected classes are removed from the initial training set. Typically, it is expected that, for most practical applications, at most only a few classes are possibly unknown. Hence, in our experiments, for the UH data set, we investigate three cases where the number of randomly removed classes ranges from 0 to 2, and for the Indian Pines data set, we remove 0 to 3 classes. When no class is removed from the initial training set, it reduces to a traditional active learning problem. A Bayes classifier is utilized for the classification part where the kernel density estimation (KDE) algorithm is used to estimate the probability density functions (pdfs) of each class by placing a Gaussian kernel around each training sample. After each active learning query, KDE updates the pdf by adding one more Gaussian kernel around the new labeled sample. Note that our framework can be applied to any classifier (e.g., support vector machine or logistic regression and their variants). However, since we require posterior probabilities to calculate entropy, in this paper, we use KDE due to its simplicity and computational efficiency.

Five active learning query strategies are implemented and compared in our experiments—random sampling (RS), BT, entropy-based uncertainty sampling (Entropy), ID, and the proposed LID. RS is implemented such that unlabeled samples are randomly selected from the candidate set $\mathcal{U}$ at each query step, labeled, and added to the training set $\mathcal{L}$.

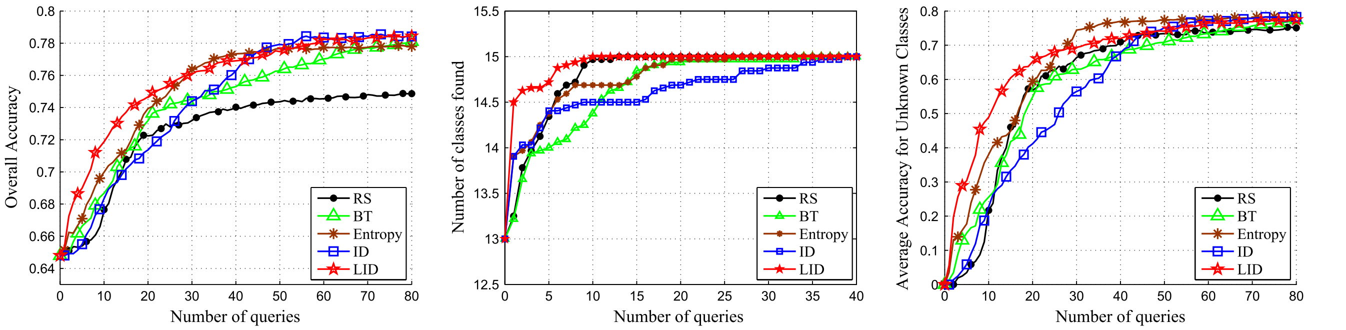
A. UH Data Set

When all the classes in the candidate set are known, the problem reduces to a traditional active learning problem. We evaluate the performance of each query strategy by constructing learning curves that plot the overall accuracies across all classes as a function of the number of queries made, which is shown in Fig. 8 where each curve is averaged across ten experiments with different randomly selected initial training sets, candidate sets, and test sets. The proposed LID-based query strategy achieves the best performance for this traditional active learning setting in the sense that it starts with a higher rate of classification improvement. Also, note that the overall accuracy for the uncertainty-based strategies (BT and Entropy) and ID drop during the first few query steps, which illustrates the point that representative samples are more important for classification



(a) (b) (c)

Fig. 9. Learning curves with one unknown class for the UH data set. (a) Overall accuracy. (b) Class discovery (number of classes found). (c) Accuracy for the unknown class.



(a) (b) (c)

Fig. 10. Learning curves with two unknown classes for the UH data set. (a) Overall accuracy. (b) Class discovery (number of classes found). (c) Average accuracy for the unknown classes.

during the initial and crucial stage of active learning where we do not have enough labeled samples.

When we have one unknown class in the candidate set, 30 experiments are implemented, and each of them has one randomly selected class removed from the starting training set $\mathcal{L}$. For the case of two unknown classes, 40 experiments are implemented with each of them having two randomly selected classes removed from the starting training set $\mathcal{L}$. For both cases, we plot the learning curves of the overall accuracies, class discovery (number of classes found), and the average accuracies for the corresponding unknown class in Figs. 9 and 10, which are computed by averaging the results across all the experiments. Again, in Figs. 9(a) and 10(a), we observe that LID achieves higher classification improvement at the beginning and reaches a higher accuracy upon convergence. More importantly, the proposed method discovers the new class much faster and significantly outperforms the other methods in classifying the new class, as shown in Figs. 9(b) and (c) and 10(b) and (c). In more detail, in the first 20 steps of active query, the LID-based method achieves an improvement of 10%–40% on the accuracy of the new class than the other baseline methods, and finally, they all converge to the same level.

As discussed in Sections II and III, both ID and LID choose entropy as the base measure in their query strategies. LID achieves improvements over entropy, but ID degrades the performance of entropy. The reason for this should be that ID only queries uncertain samples with high global density, which may not be representative for each class. On the other hand, LID queries uncertain samples with high local density, which helps the KDE-based Bayes classifier better estimate the distributions of each class and achieves better classification results.

Another phenomenon that we can observe from these figures is that uncertainty- and ID-based query strategies (BT, Entropy, and ID) are performing better than RS in overall classification for all cases in Figs. 8–10. However, as shown in Figs. 9(b) and 10(b), they are not always better than RS in detecting unknown classes due to the fact that uncertainty- and ID-based query strategies may treat the unlabeled samples of unknown classes as less informative than some samples of known classes while RS treats every unlabeled sample equally. Thus, RS can occasionally pick up the unknown classes faster than the uncertainty- and ID-based query strategies. However, the proposed LID-based approach still outperforms all other baseline methods in detecting and classifying the unknown classes.

To show the efficacy of the SELF algorithm as the dimension reduction method, another experiment with the same setup as in Fig. 9 is undertaken, except that PCA is employed as the dimension reduction method—results from this setup are provided in Fig. 11. Compared to Fig. 9 with SELF, the classification performance in Fig. 11 exhibits a drop in performance, but our proposed AL method still provides the best performance.

Since the performance of the proposed query strategy depends on the clustering quality of the DPMM, we show the

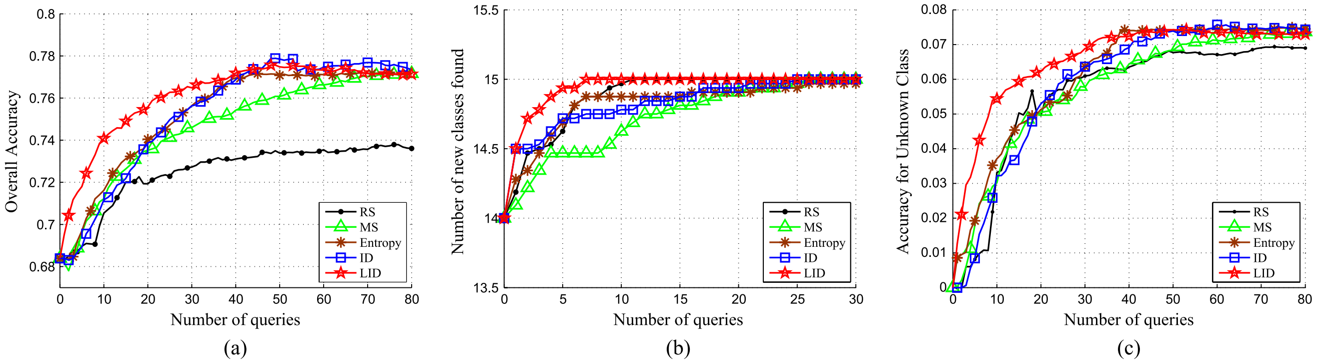


Fig. 11. Learning curves with one unknown class (with PCA as the feature reduction method) for the UH data set. (a) Overall accuracy. (b) Class discovery (number of classes found). (c) Accuracy for the unknown class.

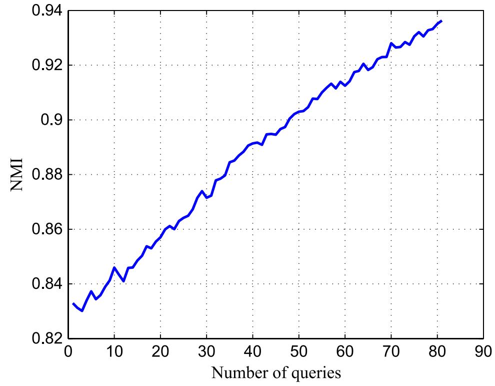


Fig. 12. Learning curves of the NMI for the cluster results given by DPMM using the UH data set with one unknown class.

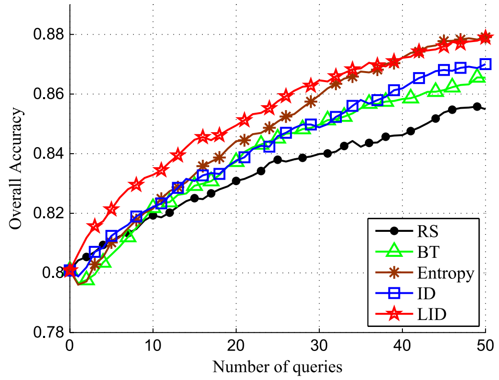


Fig. 13. Learning curves of overall accuracy for all query strategies without unknown classes for the Indian Pines data set.

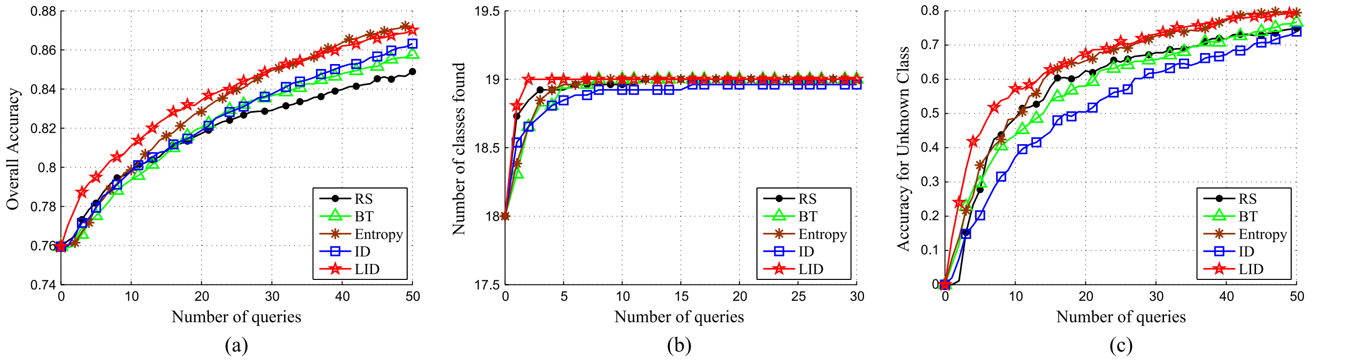
clustering quality using the normalized mutual information (NMI) [45], a standard method of measuring clustering quality, across all the active learning steps in Fig. 12. As illustrated, the clustering performance becomes better as more samples are queried because a more accurate parameter initialization and a better subspace from SELF result when the training sample size grows.

B. Indian Pines Data Set

We perform similar experiments for the Indian Pines hyperspectral data set and obtained similar results, shown in Figs. 13–16 where the number of unknown classes increases from 0 to 3. In all cases, the proposed LID-based query strategy achieved the best overall classification accuracy due to the fact that the locally representative samples that it selected are more useful for classification when we do not have enough training samples, no matter whether new classes exist or not in the unlabeled candidate pool. When new classes emerge, the proposed method will first focus on querying samples which are assigned to new clusters by DPMM and are more likely to be new classes. Thus, it significantly outperforms the other methods in discovering and classifying the unknown classes. In particular, in the first 20 steps of active query, the LID-based method achieves an improvement of about 10%–30% in the accuracy of the missing classes, compared to the other methods. We should keep in mind that the new clusters found by the DPMM do not have to be new classes, particularly in remote sensing applications. New clusters usually emerge at spatial locations which are different from the labeled data, so they can either be new classes or correspond to unrepresented areas of known classes due to the spatial variability of remote sensing images, which results in multimodal distributions of certain classes.

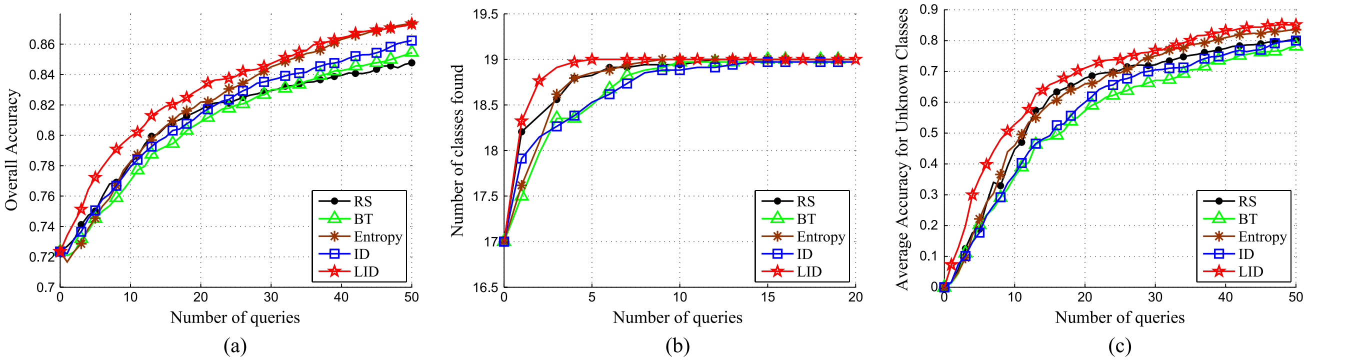
C. Visualizing the Detection of New Classes

To demonstrate the clustering performance on an image, we crop a region from the UH which contains three classes as shown in Fig. 17(a): class 2 (“Grass-stressed”), class 3 (“Grass-synthetic”), and class 15 (“Running Track”). We remove the three classes one at a time (resulting in different “unknown” classes each time). When class 15 is removed, the clustering result (including known and unknown classes) and detected new class are shown in Fig. 17(b) and (e). Similarly, results when class 3 is removed are shown in Fig. 17(c) and (f); results when class 2 is removed are shown in Fig. 17(d) and (g). As we can see in all cases, different classes are separated into different clusters [see Fig. 17(b)–(d)], and the unknown class is detected as a new cluster successfully [see Fig. 17(b)–(d)]. Note that, if there are isolated small clusters comprising very few pixels, they are likely to belong to outliers, and we ignore them.



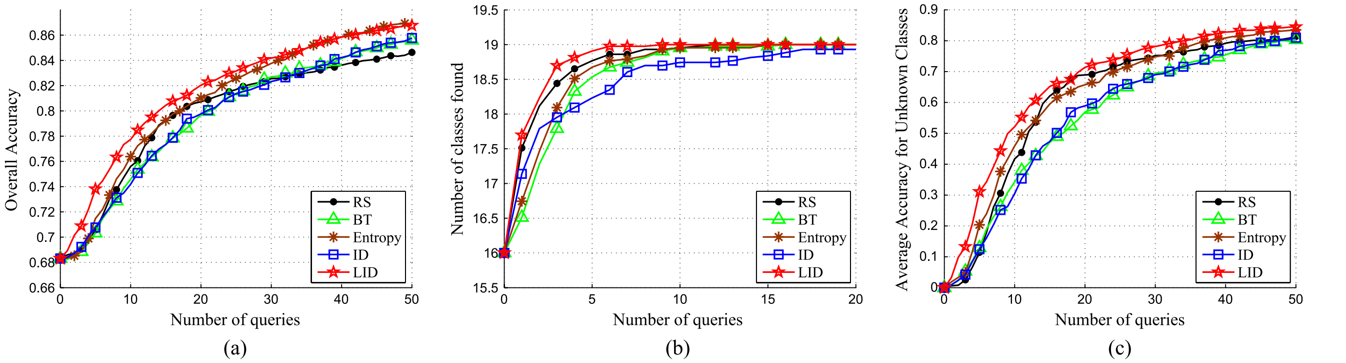
(a) (b) (c)

Fig. 14. Learning curves with one unknown class for the Indian Pines data set. (a) Overall accuracy. (b) Class discovery (number of classes found). (c) Accuracy for the unknown class.



(a) (b) (c)

Fig. 15. Learning curves with two unknown classes for the Indian Pines data set. (a) Overall accuracy. (b) Class discovery (number of classes found). (c) Average accuracy for the unknown classes.

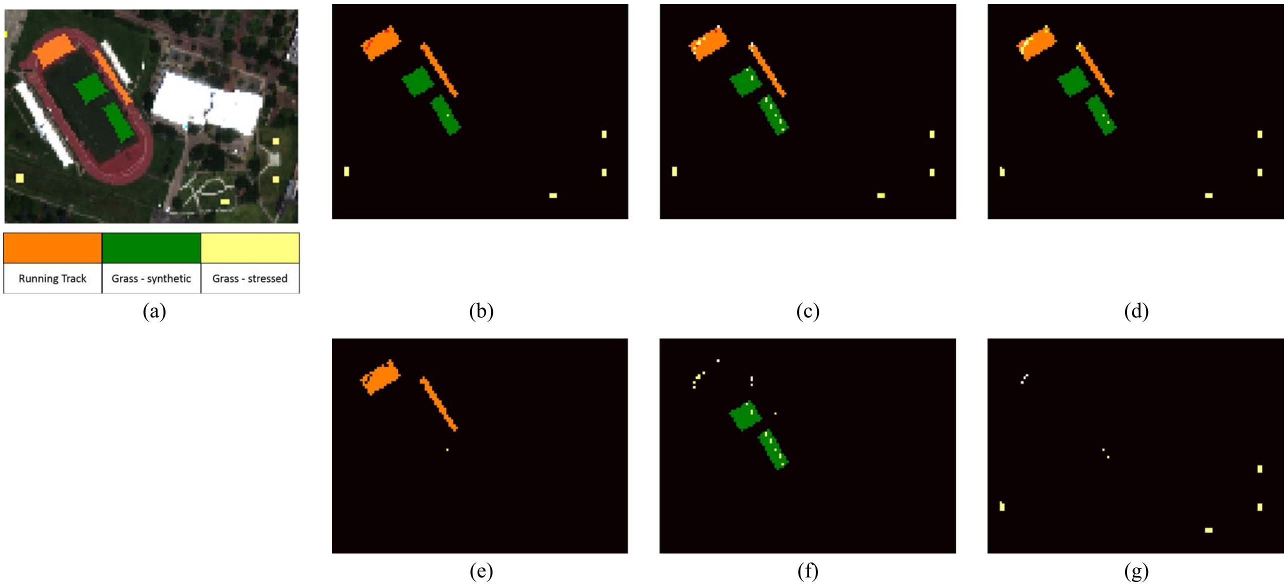


(a) (b) (c)

Fig. 16. Learning curves with three unknown classes for the Indian Pines data set. (a) Overall accuracy. (b) Class discovery (number of classes found). (c) Average accuracy for the unknown classes.

VI. Conclusion and Future Work

The new active learning paradigm introduced in this paper simultaneously improves the classification performance and discovers the unknown (missing) classes with the least effort of labeling/annotating—missing or unknown classes are commonly encountered in machine learning tasks, particularly the image analysis of remotely sensed data where it is of great value to find and label new classes and unexplored regions of existing classes (that may exhibit variability in the spectral reflectance characteristics). The framework achieves this goal by employing LID as the query strategy where the local density is obtained via DPMM-based clustering. The experimental results have shown that the proposed method provides significantly better performance than the other commonly used active learning frameworks both in the classification accuracy and detection speed of unknown classes. In ongoing work, we are exploring the development of semisupervised feature reduction algorithms that are optimized for this task. To the best of our knowledge, this is the first work that demonstrates a successful active learning paradigm that seeks out the discovery of unseen classes. This has profound benefits for a variety of applications, including the image analysis of big geospatial data cubes, where it is often not possible to have every class on the ground represented in the initial training library. Such an approach is also



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

(e) (f) (g)

Fig. 17. Visualizing the detection of new classes. (a) Ground truth over a cropped region of the UH data where there are three classes. When the first class ("Running Track") is removed, the clustering result (including known and unknown classes) and detected new class are shown in (b) and (e). Similarly, results when the second class ("Grass-synthetic") is removed are shown in (c) and (f); results when the third class ("Grass-stressed") is removed are shown in (d) and (g).

expected to be particularly advantageous to geospatial images where class variability (e.g., due to significant illumination differences such as classes under shadows) can lead to multimodal distributions which are not effectively accounted for in traditional labeled training libraries—the proposed framework can assist with enhancing the library by ensuring that sources of variability that express themselves as new clusters are systematically accounted for during the creation of a training library.

References

[1] B. Settles, "Active learning literature survey," Univ. Wisconsin-Madison, Madison, WI, USA, Comput. Sci. Tech. Rep. 1648, Jan. 2010.

[2] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 606–617, Jun. 2011.

[3] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[4] 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.

[5] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. 9th ACM Int. Conf. Multimedia*, 2001, pp. 107–118.

[6] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *Proc. IEEE ICASSP*, May 2002, vol. 4, pp. 3904–3907.

[7] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2008, pp. 1070–1079.

[8] A. McCallum and K. Nigam, "Employing EM and pool-based active learning for text classification," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 350–358.

[9] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, vol. 1, no. 2, pp. 209–230, Mar. 1973.

[10] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *J. Amer. Statist. Assoc.*, vol. 90, no. 430, pp. 577–588, Jun. 1995.

[11] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *J. Comput. Graph. Statist.*, vol. 9, no. 2, pp. 249–265, Jun. 2000.

[12] Y. W. Teh, "Dirichlet process," in *Encyclopedia of Machine Learning*. Berlin, Germany: Springer-Verlag, 2010, pp. 280–287.

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

[14] H. Wu, S. Prasad, M. Cui, N. T. Nguyen, and Z. Han, "Hyperspectral image classification based on Dirichlet process mixture models," in *Proc. IEEE IGARSS*, Jul. 2013, pp. 1043–1046.

[15] H. Wu and S. Prasad, "Infinite Gaussian mixture models for robust decision fusion of hyperspectral imagery and full waveform lidar data," in *Proc. IEEE GlobalSIP*, Dec. 2013, pp. 1025–1028.

[16] F. Wood, S. Goldwater, and M. J. Black, "A non-parametric Bayesian approach to spike sorting," in *Proc. 28th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2006, pp. 1165–1168.

[17] A. Dubey *et al.*, "Clustering protein sequence and structure space with infinite Gaussian mixture models," in *Proc. Pac. Symp. Biocomput.*, 2004, vol. 9, pp. 399–410.

[18] G. Jun and J. Ghosh, "Semisupervised learning of hyperspectral data with unknown land-cover classes," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 273–282, Jan. 2013.

[19] H. Wu, S. Prasad, and T. Priya, "Detecting new classes via infinite warped mixture models for hyperspectral image analysis," in *Proc. IEEE ICIP*, 2014, pp. 5027–5031.

[20] Y. Zhang *et al.*, "Ensemble multiple kernel active learning for classification of multisource remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 2, pp. 845–858, Feb. 2015.

[21] Y. Zhang and S. Prasad, "Locality preserving composite kernel feature extraction for multi-source geospatial image analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 3, pp. 1385–1392, Mar. 2015.

[22] M. Cui and S. Prasad, "Class dependent sparse representation classifier for robust hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2683–2695, May 2015.

[23] S. Prasad and M. Cui, "Sparse representations for classification of high dimensional multi-sensor geospatial data," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2013, pp. 811–815.

[24] M. Cui and S. Prasad, "Angular discriminant analysis for hyperspectral image classification," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1003–1015, Sep. 2015.

[25] M. Cui and S. Prasad, "Sparsity promoting dimensionality reduction for classification of high dimensional hyperspectral images," in *Proc. IEEE ICASSP*, May 2013, pp. 2154–2158.

[26] T. S. Haines and T. Xiang, "Active rare class discovery and classification using Dirichlet processes," *Int. J. Comput. Vis.*, vol. 106, no. 3, pp. 315–331, Feb. 2014.

[27] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[28] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[29] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

[30] X. He and P. Niyogi, "Locality preserving projections," *Adv. Neural Inf. Process. Syst.*, vol. 16, pp. 153–160, 2003.

[31] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[32] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, 2004.

[33] D. Lunga, S. Prasad, M. Crawford, and O. Ersoy, "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 55–66, Jan. 2014.

[34] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese, "Semi-supervised local Fisher discriminant analysis for dimensionality reduction," *Mach. Learn.*, vol. 78, no. 1/2, pp. 35–61, Jan. 2010.

[35] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proc. 11th Int. Conf. Mach. Learn.*, 1994, pp. 148–156.

[36] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. ACM 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 287–294.

[37] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1289–1296.

[38] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[39] S. Prasad, M. Cui, W. Li, and J. Fowler, "Segmented mixture-of-Gaussian classification for hyperspectral image analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 138–142, Jan. 2014.

[40] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000.

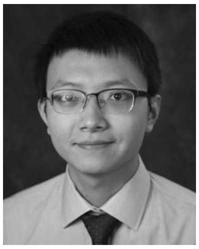
[41] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Univ. Toronto, Toronto, ON, Canada, CRG-TR-93-1, 1993.

[42] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statist. Sinica*, vol. 4, pp. 639–650, 1994.

[43] J. Pitman, "Combinatorial Stochastic Processes," Dept. Statist., Univ. California Berkeley, Berkeley, CA, USA, Tech. Rep. 621, 2002.

[44] T. Iwata, D. Duvenaud, and Z. Ghahramani, "Warped mixtures for nonparametric cluster shapes," in *Proc. 29th Conf. Uncertainty Artif. Intell.*, 2013, pp. 1–10.

[45] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.



Hao Wu (S'13) received the B.E. degree in electrical and information engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2012. He is currently working toward the Ph.D. degree in the Hyperspectral Image Analysis Laboratory within the Electrical and Computer Engineering at the University of Houston, Houston, TX, USA.

His research interests include statistical machine learning, pattern recognition, and active learning for high-dimensional image analysis.

Saurabh Prasad (S'05–M'09–SM'14) received the B.S. degree in electrical engineering from Jamia Millia Islamia, New Delhi, India, in 2003, the M.S. degree in electrical engineering from Old Dominion University, Norfolk, VA, USA, in 2005, and the Ph.D. degree in electrical engineering from Mississippi State University, Starkville, MS, USA, in 2008.

He is currently an Assistant Professor in the Electrical and Computer Engineering Department, University of Houston, Houston, TX, USA, where he leads a research group on machine learning, signal processing, and image analysis. His research interests include statistical pattern recognition, Bayesian inference, kernel methods, information fusion, and subspace learning with applications to geospatial, acoustic, and biomedical data analysis.

Dr. Prasad was awarded two research excellence awards (2007 and 2008) during his Ph.D. study at Mississippi State University, including the university-wide outstanding graduate student research award. In July 2008, he received the Best Student Paper Award at the IEEE International Geoscience and Remote Sensing Symposium 2008 held in Boston, MA, USA. In October 2010, he received the State Pride Faculty Award at Mississippi State University for his academic and research contributions. In 2014, he received the NASA New Investigator (Early Career) Award. He is an active reviewer for various journals on signal processing, image processing, and machine learning. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.