Active Learning via Multi-View and Local Proximity Co-Regularization for Hyperspectral Image Classification

Wei Di, *Student Member, IEEE*, and Melba M. Crawford, *Fellow, IEEE*

***Abstract*—A novel co-regularization framework for active learning is proposed for hyperspectral image classification. The first regularizer explores the intrinsic multi-view information embedded in the hyperspectral data. By adaptively and quantitatively measuring the disagreement level, it focuses only on samples with high uncertainty and builds a contention pool which is a small subset of the overall unlabeled data pool, thereby mitigating the computational cost. The second regularizer is based on the "consistency assumption" and designed on a spatial or the spectral based manifold space. It serves to further focus on the most informative samples within the contention pool by penalizing rapid changes in the classification function evaluated on proximally close samples in a local region. Such changes may be due to the lack of capability of the current learner to describe the unlabeled data. Incorporating manifold learning into the active learning process enforces the clustering assumption and avoids the degradation of the distance measure associated with the original high-dimensional spectral features. One spatial and two local spectral embedding methods are considered in this study, in conjunction with the support vector machine (SVM) classifier implemented with a radial basis function (RBF) kernel. Experiments show excellent performance on AVIRIS and Hyperion hyperspectral data as compared to random sampling and the state-of-the-art SVM$_{\mathrm{SIMPLE}}$.**

***Index Terms*—Active learning, classification, data regularization, hyperspectral data, multi-view learning.**

I. Introduction

Supervised classification requires labeled data, which can be costly and difficult to acquire. This problem is exacerbated by high-dimensional hyperspectral data. Active learning (AL) integrates the classifier with training set design by ranking the unlabeled data iteratively and only selecting samples with the highest training utility [7], [9], [14]. In a human–machine interaction scenario, it can provide advice to the human annotator for the next query, which aims to select the most representative samples for the chosen learner by examining the properties of the classifier through both the labeled and unlabeled data. Thus, it leads to greater information exploitation for the data and explores the maximum potential of the learner toward the given data. Also, by focusing on a much smaller but most useful sample set for the classification problem, active learning significantly reduces the cost of data collection. For existing labeled sample sets, active learning also provides capability to select the most informative subset for a given classifier.

Although active learning has been widely studied in document retrieval and natural language learning [8]–[11], [14], related work has been quite limited in remote sensing [1]–[7]. Methods employ different query strategies, such as margin sampling [2], [4], [6], uncertainty sampling [3], [4], cost sensitive sampling [5], and the query-by-committee (QBC)-based method [4], [7]. The key is to select samples with higher uncertainty or which cause greater ambiguity for the classifier.

The query-by-committee based method, which is a popular strategy in AL, utilizes the lack of consensus between a group of diverse classifier committee members [4], [7], whereby samples for which there is greatest disagreement among the committee are selected. For a discriminative classifier, another effective strategy is to select boundary samples, which often cause confusion to the classifier and are also key to building the classification hyperplane. A straightforward way for identifying boundary samples is to search for samples that are close to the current classification hyperplane [2], [4], [6], [8], e.g., margin sampling. Other strategies can also be applied, such as querying "nearby" samples that have greater inconsistency relative to the given sample by investigating local variation (such as using the local Laplacian graph [6]). Those samples often contain more uncertainty information and often lie near the hyperplane.

All of those approaches are closely related to the concept of data regularization, which has recently received much attention in machine learning. It is based on the important "consistency" assumption [18], [20], [21]. Usually, a regularization setting, which is often designed to represent the assumed smoothness on the intrinsic data structure associated with both the labeled and unlabeled points, is incorporated into the overall learning framework to implicitly or explicitly exploit the link between the marginal density $p(\mathbf{x})$ over the sample space and the conditional probability $p(y|\mathbf{x})$. The goal is to improve the conditional probability to benefit the supervised classification problem.

In this paper, we incorporate the idea of data regularization into the active learning framework and propose a novel sequential co-regularizer from two perspectives for hyperspectral image classification: 1) consistency between multiple classifiers generated from multi-view feature subsets; and 2) consistency between similar samples.

The aforementioned query-by-committee method can be regarded as seeking samples that violate the "consistency assumption." The key lies in the "value of agreement" [13]. However,

Manuscript received May 01, 2010; revised December 15, 2010; accepted December 22, 2010. Date of publication March 07, 2011; date of current version May 18, 2011. This work was conducted at the Purdue Laboratory for Applications of Remote Sensing (LARS) and supported by the National Science Foundation Grant 0705836 and LARS. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jocelyn Chanussot.

The authors are with the Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN 47907 USA (e-mail: wdi@purdue.edu; mcrawford@purdue.edu).

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

Digital Object Identifier 10.1109/JSTSP.2011.2123077

most of the QBC strategy depends on the quality of the committee which is often generated by bagging from the sample space or hypothesis space. To avoid the problem of under-representation, the number of committee members may be huge, thus resulting in the need to search and prune. Also, given that only limited labeled samples are available at the early learning stage, each subsample space may be too weak to provide a reliable representation of the data. In addition, with the high dimensionality of hyperspectral data, the "curse of dimensionality problem" might be exaggerated.

To avoid these problems, we develop the first regularizer by instead manipulating the spectral feature space to explore the intrinsic multi-view information embedded in the hyperspectral image to construct the committee. Multi-view learning, which was first proposed by de Sa [15], is shown to learn the target concept faster by exploring complementary information from disjoint sub-sets of features (views) [10], [14]. The disagreement (inconsistency) among different views bootstraps learners from each view to converge quickly to the target concept by learning from mistakes [7], [9], [10], [12]–[15]. Hyperspectral image data contain hundreds of narrow bands over an interval of the electromagnetic spectrum, providing enough complementary information to launch an effective multi-view learning strategy. Intervals of the electromagnetic spectrum differ in their discriminative ability towards classes, while naturally providing the necessary diversity for constructing the classifier committee. Generating views by segmenting the high-dimensional data in the spectral domain can mitigate the impact of small numbers of labeled samples with respect to the high dimensionality of the data. Moreover, by querying samples with higher disagreement between views, candidate samples are restricted by this regularizer within the contention pool, which is a subset of the unlabeled data. Thus, computational complexity is reduced.

Further, in order to concentrate on the most informative samples within this pool, we explore sample consistency by the clustering assumption [20], [21]:

> *Similar samples (or points on the same data structure such as the cluster or manifold) are likely to have the same label; and samples which lie in the low density region of a class, where the classification boundary may cross, also tend to show greater inconsistency toward nearby samples.*

Querying those samples could help refine the classification hyperplane, especially for discriminative classifiers such as SVM [35], whose performance heavily relies on the quality of the support vectors around the decision boundary. Also, by avoiding queries of samples from high-density regions where labels are more likely to be consistent, we can avoid inclusion of non-informative/redundant samples into the training pool, thereby reducing the overall sample size required to train a good learner.

Two types of intrinsic data structure spaces are investigated in this study: the image spatial space and the low-dimensional spectral manifold space. High-dimensional hyperspectral data usually lie on certain low-dimensional manifold structures [22], [23]. Remote sensing data of a given class typically occur in spatially contiguous clusters; thus, the image spatial space can be viewed as the most natural "low-dimensional manifold space." If points exhibit less consistency toward spatially neighboring samples, the current learner may lack ability to correctly discriminate those data.

Further, in the spectral domain, we apply manifold learning to search for intrinsic low-dimensional structures of hyperspectral data [19], [22]–[27]. The high-dimensional data are parameterized by seeking a smooth low-dimensional surface, whereby local pairwise distances are preserved, and similar samples are moved into closer proximity. This reinforces the "clustering assumption" [18], [20], [21] and improves the quality of the distance measure, which is key to identifying "nearby" samples [30], [31]. Because we focus on local structure, we investigate two local manifold learning methods which have performed well in our previous studies: locally linear embedding (LLE) [24] and local tangent space alignment (LTSA) [25].

Finally, a co-regularizer, which jointly combines view-disagreement and local inconsistency, is developed. The concept of "locality" in the second regularizer is defined on a spatial or spectral manifold space, which seeks to represent the intrinsic structure of hyperspectral image data from the spatial or spectral perspectives, respectively.

The remainder of the paper is organized as follows. The active learning framework based on multi-view disagreement and local proximity data regularization is presented in Section II. Section III illustrates the multi-view disagreement based regularizer; the local proximity regularizer, which is based on the spatial/spectral manifold space, is described in Section IV. Experiments and analysis are presented in Section V, and a summary is provided in Section VI.

II. Data Regularization Based Active Learing

Denote each sample $\mathbf{x} \in R^{\mathrm{B}}$ as drawn from an instance space X, and Y as the label set $\{\omega_1, \omega_2, \ldots \omega_{N\mathrm{c}}\}$ containing N_{c} classes. B is the dimension of the sample space. The purpose of classification is to learn a hypothesis h: $X \rightarrow Y$ to find the correct label $\hat{y} \in Y$

$$\hat{y} = f(\mathbf{x}, y, \{D_L, D_U\}) \tag{1}$$

where f is the classification function under hypothesis h. D_L is the labeled data pool that contains N_L samples, and D_U data is the unlabeled data pool with N_U samples, where in general $N_L \ll N_U$. In the transductive framework, the sample to be classified is from D_U, and in the inductive case, it is from the unseen data set $\mathrm{D}_{\mathrm{T}} = (\mathrm{D}_L \cup \mathrm{D}_U)^c$. Incorporating information from the unlabeled data in learning f can often lead to better generalization ability of the trained classifier [15], [21].

Consistency, which implies that the label of a data point can potentially be well estimated based on its neighbors, is commonly assumed for learning a classification mapping [18], [20], [21]. Under this assumption, the classification function can be obtained by reinforcing a regularizer $\mathfrak{R}(f, \{\mathrm{D}_U, \mathrm{D}_L\})$, which usually penalizes lack of smoothness of the classification function evaluated by both the labeled and unlabeled samples.

We adopt this idea into the active learning framework and define the loss at the τth query to reflect the overall inconsistency, i.e., the degree to which the current classifier violates the consistency assumption evaluated by all the unlabeled data $\mathbf{x}_{j,U} \in D_U^\tau$

$$Loss(\tau) = \frac{1}{N_U^\tau} \sum_{j=1}^{N_U^\tau} \mathfrak{R}(f^\tau, \mathbf{x}_{j,U}). \tag{2}$$

In AL, our purpose is to improve the learner by selecting n_Q new samples $\left\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{n_Q}\right\} \in D_U^\tau$ for query each time to maximally reduce the loss

$$\arg \max_{\left\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{n_Q}\right\} \in D_U^\tau} Loss\,(\tau - 1) - Loss\,(\tau). \tag{3}$$

Note that, with $n_Q > 1$ (batch-mode learning), an additional diversity measurement should be applied to select the n_Q best candidates to avoid inducing redundancy into the training set. This criterion is equivalent to querying the next most diverse n_Q samples that satisfy

$$\mathfrak{R}\left(\mathbf{q}_1\right) \geq \mathfrak{R}\left(\mathbf{q}_2\right) \geq \cdots \geq \mathfrak{R}\left(\mathbf{q}_{n_Q}\right) \geq \mathfrak{R}\left(\mathbf{x}'\right) \tag{4}$$

where $\mathbf{x}' \in \left(S_Q^\tau\right)^c$ is the complement set of $S_Q^\tau = \left\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{n_Q}\right\}$ with $S_Q^\tau \cup \left(S_Q^\tau\right)^c = D_U^\tau$.

At each iteration, the classifier is updated by

$$f^{\tau+1} \leftarrow \arg\min_f \left(\frac{1}{N_L^\tau} \sum_{i=1}^{N_L^\tau} V(\mathbf{x}_i, y_i, f) + \frac{1}{n_Q} \sum_{j=1}^{n_Q} V(\mathbf{q}_j, y_j, f) \right)_{\mathbf{q}_{j \in S_Q^\tau}} \tag{5}$$

where V is the loss in the classification metric. In real applications with noisy data, sampling data with the highest inconsistency value measured by $\mathfrak{R}$ can result in noise or outliers being introduced into the training set. Thus, we introduce a relaxation variable α, and first query $\alpha \times N_U \geq n_Q$ samples from the current unlabeled data pool according to (4), and then further randomly select n_Q samples from this subset.

The design of the regularizer is key to the success of the active learning strategy. According to the consistency assumption, it should favor the changes of $p(y|\mathbf{x})$ in regions with lower values of $p(\mathbf{x})$, where the decision boundary may be located. Samples which are in close proximity, but violate the consistency assumption, i.e., similar samples with higher confliction in terms of the conditional probability $p(y|\mathbf{x}_i)$ and $p(y|\mathbf{x}_j)$ should be queried first. Also, in the AL scenario, the information should be incorporated from both the labeled and unlabeled data, as well as the chosen classifier. Thus, we propose the following co-regularizer

$$\mathfrak{R}(\mathbf{x}) = \mathfrak{R}_{\mathrm{AMD}} \mathfrak{R}_{\mathrm{LIC}}(\mathbf{x}), \mathbf{x} \in \mathrm{D}_U. \tag{6}$$

The first factor is the multi-view adaptive maximum disagreement (MV-AMD) regularizer $\mathfrak{R}_{\mathrm{AMD}}$, and the second is the local inconsistency (LIC) regularizer $\mathfrak{R}_{\mathrm{LIC}}$ which represents the lack of smoothness measured on a local graph in the manifold space. Both are defined in Sections III and IV, respectively.

III. Multi-View Disagreement Based Regularization

A. Multi-View Generation for Hyperspectral Image Data

In a single-view scenario, a learner can access the entire set of domain features. In the multi-view setting where there are N_v views, the available attributes are decomposed into disjoint sets $X^1 \times X^2 \times \cdots X^{Nv}$. An instance $\mathbf{x}$ is therefore viewed as $(\mathbf{x}^1 \times \mathbf{x}^2 \times \cdots \mathbf{x}^{Nv})$. It is assumed that each view is sufficient to learn the target concept (compatibility), which means that the hypothesis from any view h_i: $X^i \rightarrow Y$ corresponds to the target hypothesis h. Learning is conducted by utilizing the complementary information between views [10]. It has been shown that minimizing the disagreement between the outputs from individual views is a sensible approximation to minimizing the misclassifications in each view [7], [10], which ultimately forces the learner group to learn the correct concept faster.

Two basic requirements for view generation are compatibility and independence [10]. Compatibility ensures that learners can ultimately converge to the same target concept, while independence is key to generating disagreement information to bootstrap each view. However, it has been shown that both are too strong in real applications, and can be relaxed without sacrificing the learning efficiency [17]. Muslea *et al.* [17] showed that the active learner can still be effective when the compatibility assumption is violated. By querying the true labels, unlike situations in semi-supervised multi-view learning [21], AL has more stable convergence towards the target concept and can compensate for correlation between views. It is shown that even with weak correlation, the ratio of the number of contention points (samples for which the evaluations from different views disagree) to the unlabeled samples still represents an upper-bound on the learning error for pair-wise views [12], [17].

To exploit the "value of disagreement" [13], diversity is a key for view generation. This guarantees that additional information can be provided by the other views to improve the learner, and it is unlikely that the learners from different views agree on an incorrect result [10], [13]. In our case, each view is obtained by the subspace grouping method [28], [29], whereby the hyperspectral data cube is segmented naturally into several disjoint contiguous sub-band sets along the spectral dimension according to the band correlation index. Each subspace has highly correlated members, but low correlation with other sub-band sets. Due to the intrinsically different spectral information contained in various spectral ranges of the data, diversity can be satisfied with less redundancy (e.g., lower correlation) between views. Further, since it is unnecessary to have label information to compute the band correlation coefficients, both the labeled and unlabeled data can be used, which ultimately improves the generalization of the classifier.

Fig. 1 shows the correlation coefficient matrix generated by the data used in the experiments [KSC (Kennedy Space Center data)], where the brighter red color denotes higher correlation, and the blue color denotes lower correlation. Blocks along the diagonal of the matrix are detected by simple edge detection and then used to generate different views.

B. Adaptive Maximum Disagreement Regularizer (AMD)

First, we define the disagreement between classification functions from all views as the associated differences in the predicted labels evaluated by unlabeled data

$$d\left(f_v^1, f_v^2, \ldots, f_v^{N_v}\right) \triangleq \frac{1}{N_U} \sum_{t=1}^{N_U} \left(\sum_{i=1}^{N_v} \sum_{j=i+1}^{N_v} 1_{\left(f_v^i\left(\mathbf{x}_t^i\right) \neq f_v^j\left(\mathbf{x}_t^j\right)\right)} \right) \tag{7}$$

where f_v^i is the classification function under hypothesis h_i for view i. The value of $d\left(f_v^1, f_v^2, \ldots, f_v^{N_v}\right)$ indicates the distance

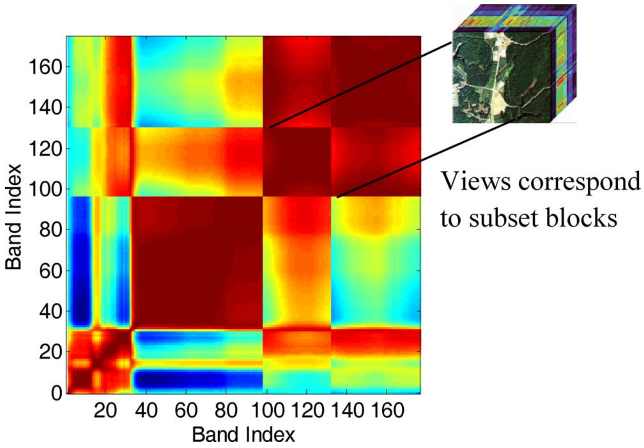


Fig. 1. Correlation coefficient matrix of KSC data.

of current multi-view learner group towards the target hypothesis. To evaluate the contribution of each sample $\mathbf{x} \in D_U$, it is decomposed into a sample based distance measure as:

$$d\left(\mathbf{x}, f_v^1, f_v^2, \ldots, f_v^{N_v}\right) \triangleq \sum_{i=1}^{N_v} \sum_{j=i+1}^{N_v} 1_{\left(f_v^i(\mathbf{x}^i) \neq f_v^j(\mathbf{x}^j)\right)}. \tag{8}$$

This sample-wise distance represents individual uncertainty as it contributes to the overall confusion. To incorporate global information from the entire unlabeled data pool, we define the maximum disagreement $MaxD(\mathrm{D}_U)$ as

$$MaxD\left(D_U\right) = \max_{\mathbf{x} \in D_U} d\left(\mathbf{x}, f_v^1, f_v^2, \ldots, f_v^{N_v}\right). \tag{9}$$

Inconsistency by different views is then defined as

$$\varphi\left(1 - \frac{d(\mathbf{x}, f_v^1, f_v^2, \ldots, f_v^{N_v})}{MaxD}\right) \tag{10}$$

where φ is a monotonic decreasing function defined on $[0, +\infty)$ with more emphasis on samples with higher disagreement level, such as the log function or hat-like loss function. To concentrate on the most informative samples, we use the indicator function to build the multi-view adaptive maximum disagreement (AMD) regularizer

$$\mathfrak{R}_{\mathrm{AMD}} = \delta\left(1 - \frac{d\left(\mathbf{x}, f_v^1, f_v^2, \ldots, f_v^{N_v}\right)}{MaxD}\right). \tag{11}$$

By using the indicator function, only samples with the maximum level of disagreement are selected into the first stage contention pool C_{AMD}, thereby restricting the query candidates to a smaller subset of the unlabeled data and reducing the computational cost. Samples in C_{AMD} represent the maximum disagreement from all views, and thus contain the most uncertainty information for the learner. Querying samples from C_{AMD} will bootstrap views to best learn the training set, and to "agree" with each other on the extra unlabeled samples [13]. Also, according to the compatibility assumption, the regularizer actually penalizes changes in the probability $p(y|\mathbf{x}^j)$, $j = 1, 2, \ldots, N_v$. We denote this active learning method by randomly sampling points from C_{AMD} as multi-view AMD-SR (Multi-view Adaptive Maximum Disagreement Single Regularization active learning). As learning progresses, different views tend to agree with each other, and the confliction level decreases, possibly resulting in inflation of the contention pool. Thus, based on the consistency assumption, we further apply the local consistency regularizer to evaluate the samples in C_{AMD} so as to force the learner to focus on querying the most informative samples.

IV. Regularization via a Generalized Manifold Space

A. Generalized Manifold Space

According to the consistency assumption [18], [20], a similarity function should be first defined to measure the closeness of samples in the input space in a meaningful way. Hyperspectral data have two important features: spatial and spectral information. Manifold learning has been demonstrated to yield higher classification accuracies and improved representation of phenomena relative to linear dimensionality reduction methods in several remote sensing investigations [22]–[27]. Hyperspectral data lie on low-dimensional manifolds that are often inherently nonlinear due to the scattering in the atmosphere and within the ground resolution cell. The resulting low-dimensional manifold coordinate system enforces the clustering assumption, whereby similar samples are moved into closer proximity, and distances between non-neighbor samples are increased. Since we focus on the local proximity property, only local manifold methods are used in our experiments: i.e., LLE [24] and LTSA [25]. Both methods start with finding the k-nearest neighbors (based on Euclidean distance) for each sample. LLE assumes that the embedding mapping is locally linear and uses constrained optimization formulation to find the optimum local convex representations of each point from a linear composition of its neighbors. LTSA estimates the local tangent spaces for each point by performing PCA on its neighborhood set, and then aligns those tangent spaces to find the global coordinates by minimizing a cost function that allows any linear transformation of each local coordinate. This ultimately results in an eigenvalue problem, whereas eigenvectors corresponding to the 2nd to $d+1$ smallest eigenvalues of the constructed alignment matrix are found as the global coordinates in the embedding space, where d is the dimension of the new generated manifold space, and $d \leq B$.

Analogous to spectral manifolds, spatial content that corresponds to the location of a point in the remotely sensed image can be also viewed as a natural manifold space parameterized by a 2-D coordinate system. This spatial manifold assumes that natural spatial clustering exists in the image, which is appropriate for many natural landscapes where classes form contiguous patches distributed across the scene.

B. Local Proximity Based Regularizer

Because global behavior of the similarity function is not critical to the clustering assumption, we restrict the "close" samples within a local k-nearest neighborhood in the low dimensional spatial or spectral manifold space. Each sample in the derived manifold space $\mathcal{M}$ is denoted as $\mathbf{z} \in R^{\mathrm{d}}$. The k-nearest neighbor set $N_{k,i}$ for a sample $\mathbf{z}_i \in D_U^* \cap C_{\mathrm{AMD}}^*$ is defined as

$$\begin{aligned} N_{k,i}(\mathbf{z}_i) = \{&\mathbf{z}_j \in D_L^* \cup D_U^* \,|\mathrm{dist}(\mathbf{z}_j, \mathbf{z}_i) \leq \mathrm{dist}(\mathbf{z}, \mathbf{z}_i)\} \\ &\text{s.t. } \mathbf{z} \neq \mathbf{z}_j, \ \forall \mathbf{z} \in D_L^* \cup D_U^*, \ \|N_{k,i}\| = k \end{aligned} \tag{12}$$

where D_L^* and D_U^* are the labeled and unlabeled data sets in the manifold space $\mathcal{M}$ corresponding to D_L and D_U, respectively. C_{AMD}^* is the contention pool in $\mathcal{M}$ corresponding to C_{AMD}. The geodesic distance corresponds to the Euclidean distance in manifold space; thus we simply use $dist(\cdot)$ as the Euclidean distance between two vectors.

A k-nearest neighborhood graph for each unlabeled sample is then defined upon its closest k samples. The graph $G(N_{k,i}, E_{k,i})$ contains vertices (nodes) and edges where $N_{k,i}$ represents the set of nodes, and $E_{k,i}$ represents the edges of the graph. The predicted labels of the unlabeled data and the true labels from the labeled data are all represented on the graph. Let L_{ij} be the length of the edge from node i to node j, which represents the inconsistency between these two nodes

$$L_{ij} = w(\mathbf{x}_j)\|\mathbf{z}_i - \mathbf{z}_j\|_2 \left(1 - \delta(f(\mathbf{x}_i) - y_j')\right) \quad (13)$$

where

$$w(\mathbf{x}_j) = \begin{cases} w_L, & \mathbf{x}_j \in D_L \\ w_U, & \mathbf{x}_j \in D_U \end{cases}, y_j' = \begin{cases} y_j, & \mathbf{x}_j \in D_L \\ f(\mathbf{x}_j), & \mathbf{x}_j \in D_U \end{cases}$$

and $w(\mathbf{x}_j)$ is the weight used to differentiate the confidence assigned for the true label and the estimated label from the training data and the unlabeled data, respectively; generally $w_L > w_U$. The length is actually a function of the conditional probability $p(y|\mathbf{x})$ and the marginal probability density $p(\mathbf{x})$. In semi-supervised learning, a regularizer should penalize changes in $p(y|\mathbf{x})$ more in the regions where values of $p(\mathbf{x})$ are smaller. Contrary to this principle, samples that lie in the lower $p(\mathbf{x})$ region, but have greater changes in $p(y|\mathbf{x})$ are of the most interest in active learning. $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ represents the information in $p(\mathbf{x})$. Intuitively, if samples lie in a lower $p(\mathbf{x})$ region, the distance between the core node and its k-nearest samples is larger than that in the high-density region. A better classifier can be obtained by querying samples from those low $p(\mathbf{x})$ regions. The overall local inconsistency score is defined as the sum of all the edge lengths in the local graph

$$\mathfrak{R}_{\mathrm{LIC}}(\mathbf{x}_i) = \sum_{\mathbf{z}_j \in N_{k,i}} L_{ij} \quad (14)$$

where $\mathbf{x}_i \in D_U \cap C_{\mathrm{AMD}}$. $\mathfrak{R}_{\mathrm{LIC}}$ represents the inconsistency of the core node (unlabeled sample) towards its neighbors evaluated on the weighted k-nearest neighborhood local graph. It also discriminatively interpolates the information from both the labeled and unlabeled samples. By using the hard label estimation y_j', the method aims to directly improve the conditional relationship using the empirical error.

Finally, the regularizer for each unlabeled sample $\mathbf{x}$ is obtained as the product of the disagreement based regularizer $\mathfrak{R}_{\mathrm{AMD}}(\mathbf{x})$ and the local inconsistency regularizer $\mathfrak{R}_{\mathrm{LIC}}(\mathbf{x})$ in (6). Three methods are developed under this framework according to different "manifold spaces" that are used:

- SpaCR: Spatial manifold space $\mathcal{M}_{\mathrm{Spa}}$ based co-regularization AL (Spatial-CR);
- LmCR: LLE manifold space $\mathcal{M}_{\mathrm{LLE}}$ based co-regularization AL (LLE-mCR);
- TmCR: LTSA manifold space $\mathcal{M}_{\mathrm{LTSA}}$ based co-regularization AL (LTSA-mCR).

To evaluate the effectiveness of using the low-dimensional manifold space, we denote SpeCR (Spectral feature space based co-regularization AL: Spectral-CR) as the method which employs the two-stage co-regularizer where the second local inconsistency regularizer uses the original high-dimensional spectral features to search the k-nearest close samples to build the graph and compute the distance between $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ in (13).

V. Experiments

A. Data Description

Two hyperspectral data sets from different sensors are used for this experiment [3], [32], [33]. NASA EO-1 Hyperion and Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), were used to collect the data over the Okavango Delta, Botswana (BOT), and the Kennedy Space Center (KSC), respectively. The KSC data were acquired in March 1996 and consist of 224 bands of 10-nm width covering 400–2500 nm, with 18-m spatial resolution. Discrimination of the land cover types in the KSC data is difficult due to the similarity of the spectral signatures for certain vegetation types and the existence of mixed classes [3]. BOT data were acquired in May 2001 in a 7.7-km strip at 30-m spatial resolution with 242 bands of 10-nm width covering 400–2500 nm. The data were obtained to study the impact of flooding on vegetational response. Removal of noisy and water absorption bands resulted in 176 and 145 candidate features for KSC and BOT data, respectively. Details of the land cover classes of both data sets are given in Table I. Figs. 2 and 3 contain the RGB images of a portion of the whole scene and the corresponding distribution of the labeled data by class for the two data sets. Classes occur in small patches scattered throughout the image.

B. Experimental Design

The labeled samples from each data set were randomly sampled into two equal sets: one for transductive learning, which was used to generate the initial labeled training set (D_L) and the unlabeled data set (D_U); the other was for inductive learning and used as the unseen data set (D_{T}). Five views were generated for both data sets based on the correlation matrix. For the KSC data, the associated band indices are 1–11, 12–31, 32–96, 97–130, and 131–176; and for the BOT data, views correspond to bands 1–25, 26–61, 62–79, 80–110, and 111–145. The initial labeled data pool for KSC consists of only 30 pixels obtained by randomly selecting three samples from each class. For the BOT data set, the initial labeled data pool consists of 54 pixels obtained by randomly choosing six samples from each class. A larger initial set was used for the BOT data because of its lower spatial resolution and lower SNR as compared to the AVIRIS data. All the initial sets are quite small relative to the dimension of the hyperspectral image. To concentrate on the early performance of active learning, algorithms ran for 400 and 600 epochs for KSC and BOT data, respectively, adding one pixel to D_L at each iteration (details are listed in Table II).

The focus of this research is active learning, so only one method was used for classification. Support vector machines (SVMs) are nonparametric discriminative classifiers which typically perform well in classification of hyperspectral data and do not require reduction in dimensionality [35]. A radial basis function (RBF) kernel-based SVM [34] was used as the base

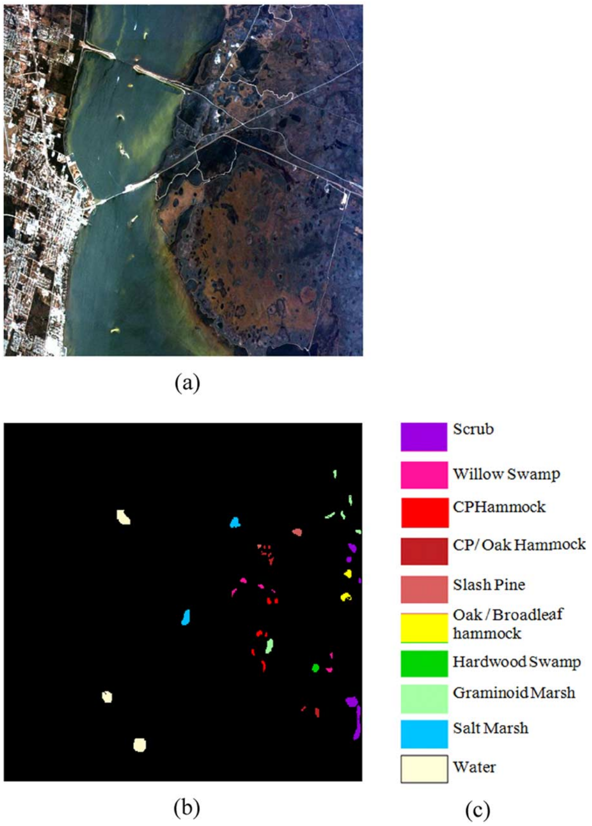


Fig. 2. (a) RGB image of KSC data. (b) KSC labeled data. (c) Class legend.

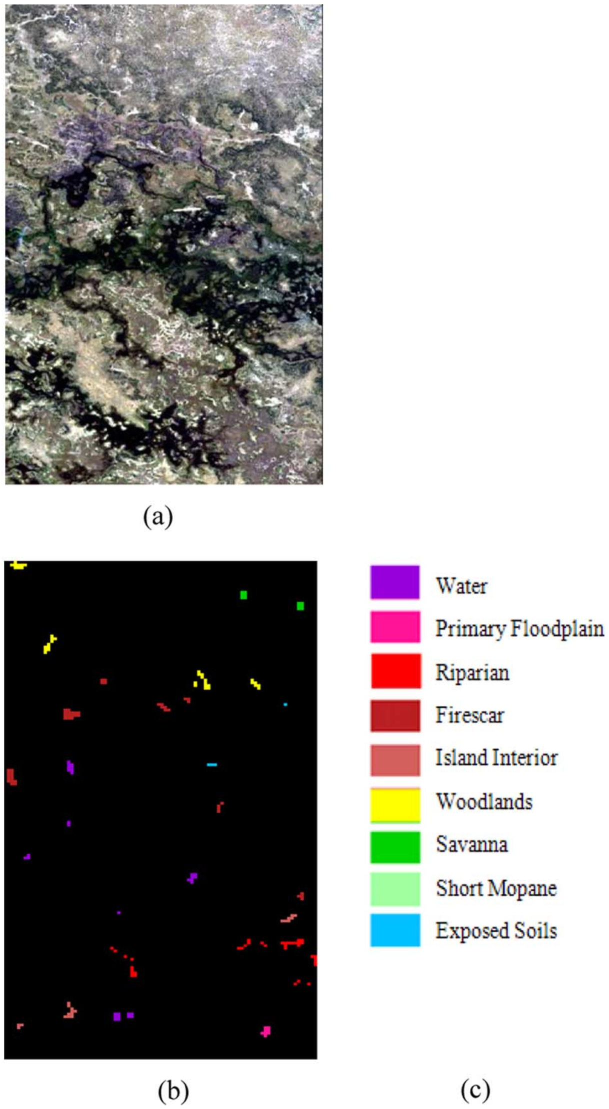


Fig. 3. (a) RGB subset of BOT data. (b) BOT labeled data. (c) Class legend.

TABLE I
CLASS INFORMATION FOR KSC AND BOT DATA

	Kenney Space Center (KSC)		Botswana (BOT)	
	Class Name	No. of Samples	Class Name	No. of Samples
1	Scrub	761	Water	297
2	Willow Swamp	243	Primary Floodplain	437
3	Cabbage Palm Hammock	256	Riparian	448
4	Cabbage Palm/Oak Hammock	252	Firescar	354
5	Slash Pine	161	Island Interior	337
6	Oak / Broadleaf hammock	229	Woodlands	357
7	Hardwood Swamp	105	Savanna	330
8	Graminoid Marsh	431	Short Mopane	239
9	Salt Marsh	419	Exposed Soils	215
10	Water	927		
	Total Number Samples	3784		3014

learner. All data were first normalized [34] to avoid scaling issues in computing the kernel. Two hyperparameters, g (spread of the Gaussian kernel function) and C (regularization parameter), were obtained by grid search over a wide range of values. In order to focus on construction of the training data set by AL, parameters were selected such that satisfactory performance was achieved when using all the available training data ($\mathrm{D}_L \cup \mathrm{D}_U$) which occurs at the later learning stage. Empirically, we found that a wide range of parameter values perform well for these data sets. Parameters were not updated along the learning since no significant improvement was observed in preliminary experiments. Classification results were obtained by training the base learner on the final labeled data pool using the full set of spectral data. Results are reported by the average performance of at least ten-fold cross-validation experiments.

To evaluate the approach, we compare the manifold based co-regularization methods: SpaCR, LmCR, and TmCR with 1) SpeCR to show the incremental contribution of nonlinear dimensionality reduction), 2) AMD-SR to evaluate the incremental contribution by the additional local inconsistency regularizer, and 3) the base-line random sampling (RS) and the benchmark AL method: SVM-based simple “Margin Sampling” [11] (denoted by $\mathrm{SVM}_{\mathrm{MS}}$). To directly evaluate the overall performance compared to passive learning (RS), we define the average improvement of the classification accuracy (D) as

$$D = \frac{1}{N_q} \sum_{i=1}^{N_q} (Acc_{AL_i} - Acc_{RS_i}) \tag{15}$$

and the Efficiency Ratio (ER) as

$$ER = \frac{\sum_{i=1}^{N_q} (Acc_{AL_i} - Acc_{RS_i})}{\sum_{i=1}^{N_q} (\max_{1 \leq i \leq N_q} Acc_{AL_i} - Acc_{AL_i})} \tag{16}$$

TABLE II
SAMPLE CHARACTERISTICS FOR EXPERIMENTS

Data Set		KSC	BOT
Total No. of Samples		3784	3014
No. of Classes		10	9
Band Indices of Each View		1-11, 12-31, 32-96, 97-130, 131-176	1-25, 26-61, 62-79, 80-110, 111- 145
Transductive Learning	Total No. of Samples	1892	1507
	Initial No. of Samples of Each Class in D_L	3	6
	Initial Size of D_L	30	54
	No. of Queries (N_q)	400	600
	Final Size of D_L	430	654
	Final Size of D_U	1462	853
Inductive learning	Size of D_T	1892	1507

TABLE III
ABBREVIATIONS AND RELATED PARAMETER SETTINGS FOR EXPERIMENTS

	KSC		BOT	
AMD	5 views			
SpeCR SpaCR LmCR TmCR	d= 2 (for SpaCR), 8, 15 k = 5, 10, 15, 20 ρ=10, 20, 30 where $k \leq \rho$	SpaCR	k=10	d = 2
		LmCR$_{\mathrm{I}}$		d = 8, ρ =20
		LmCR$_{\mathrm{II}}$		d=15, ρ =15
		TmCR$_{\mathrm{I}}$		d=8, ρ= 20
		TmCR$_{\mathrm{II}}$		d=15, ρ =15

Notes: α = 0.1; d = dimension of derived manifold space dimension, ρ= local nearest neighborhood size for manifold learning, k = k-nearest neighborhood size in Eq.(12) for computing the local regularizer.

where N_q is the total number of query steps. Acc_{AL_i} and Acc_{RS_i} represent the classification accuracy by a given active learning method and random sampling at the ith query, respectively. Parameter settings for the experiments and the corresponding abbreviations are listed in Table III. Three parameters are required: 1) the k-nearest neighborhood size for manifold learning, denoted as ρ; 2) the dimension of the generated manifold space, denoted as d; 3) the k-nearest neighborhood size used to compute the local inconsistency, denoted as k.

Note that, in SpeCR and SpaCR, only one parameter (k) is needed, while in LmCR and TmCR all three parameters are required. The relaxation variable that mitigates the effect of noise and outliers was set as 0.1 in our experiments, and values of $w_L = 3$ and $w_U = 1$ were assumed. The absolute values of w_L and w_U are not very important since they only work to enlarge the discrepancy between the confidence assigned to the labeled and unlabeled data, and the query is based only on the ranking order of the local inconsistency. The parameter ρ used in manifold learning is the number of the nearby samples that are assumed to have a strong local geometric relationship. It constrains the size of the local region that is used to construct the spectral graph for generating the manifold coordinates. Thus, we only tested the cases with the k-nearest neighborhood size $k \leq \rho$. Also, it should be noted that in LTSA, ρ must be larger than the desired dimension of the generated manifold space since each local region with size ρ is used to perform a local PCA-like spectral analysis, and then aligned to obtain the global manifold coordinate.

For the KSC data, all methods (SpeCR, SpaCR, LmCR, and TmCR) were tested with various sizes of k-nearest neighborhoods: k =5, 10, 15, and 20, respectively. LmCR and TmCR were evaluated with d values of 8 and 15 and ρ =10, 20, and 30. Only $k = 10$ was tested for the BOT data based on our previous studies using these data [26], [27].

C. fResults and Discussion

Fig. 4 shows some typical examples of classification accuracy of the proposed methods compared with RS and $\mathrm{SVM_{MS}}$ for D_U and $\mathrm{D_T}$ from both data sets. Parameter settings are: KSC with $k = 15$, $d = 15$, $\rho = 30$, and BOT with $k = 10$, $d = 8$, $\rho = 20$. Fig. 5 and Table IV provide comparisons in terms of Efficiency Ratio (*ER*) and the incremental change in classification accuracy (D) between different methods (given in Table III) for BOT data. Table V lists results of different methods and settings for the KSC data in terms of the average improvement in classification accuracy (D).

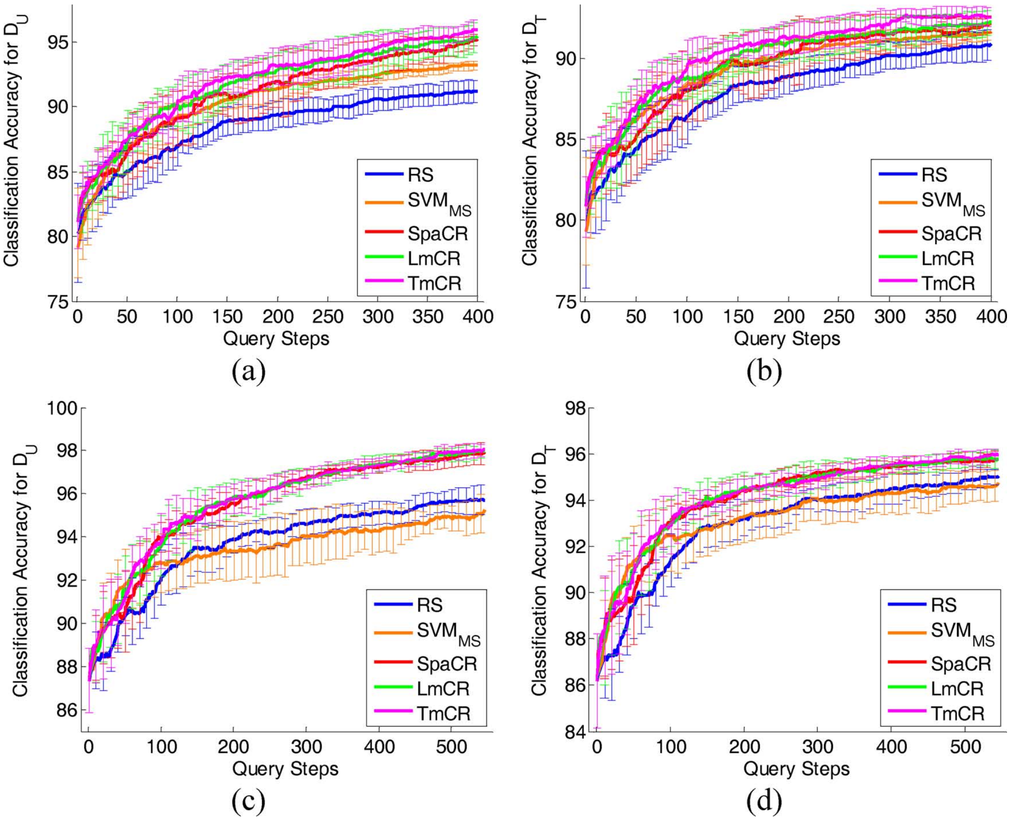
The following trends are observed.

The proposed active learning methods, as well as AMD and SpeCR, outperform random sampling and $\mathrm{SVM_{MS}}$ in all cases for the two data sets, based on overall classification accuracies (Tables IV and V) and the Efficiency Ratio (Fig. 5). Classification accuracies by active learning (including $\mathrm{SVM_{MS}}$) on D_U and $\mathrm{D_T}$ are all higher than for RS, but the difference is smaller for unseen data. This is because information from the unlabeled data is incorporated directly into the classifier during the learning stage as samples from D_U are iteratively queried and evaluated.

Several specific aspects of the proposed methods were also evaluated, including: the co-regularization versus AMD based single-regularization; spatial versus spectral local regularization; and the contribution of individual views.

1) Comparison of AMD With Co-Regularization Methods: Table V shows that the proposed co-regularization methods (SpaCR, LmCR, and TmCR) have better performance than multi-view based AMD single-regularization in terms of incremental classification accuracy for both D_U and $\mathrm{D_T}$. This indicates the potential value of adding the second regularizer to the active learning process. Table V also indicates that results for the proposed methods are insensitive to the values of k for these data. The spectral based method SpeCR also outperforms AMD in all cases on D_U, and all but two cases ($k = 5$, $k = 20$) on the unseen data. SpeCR is more affected by k than the other proposed methods, with the best performance being achieved when $k = 10$.

2) Evaluation of Co-Regularization Methods of Different Manifold Spaces: The two local manifold learning based methods (LmCR and TmCR) produced similar results. Both methods have better performance than SpeCR in most cases, and the best performance is obtained by TmCR ($d = 15$, $\rho = 30$, and $k = 15$). This is consistent with our previous study in which we found that LTSA yielded higher accuracies on the KSC data [26], [27]. These results indicate that manifold learning successfully captures spectral characteristics of the data. Moreover, by reducing the dimension from the original spectral feature space (for KSC, $B = 176$ and for BOT, $B = 145$) to the lower dimensional manifold space (d =8 and 15), which can be computed offline, LmCR and TmCR greatly



(a) (b)

(c) (d)

Fig. 4. Classification accuracies for KSC (a) D_U, (b) D_T, and for BOT: (c) D_U, (d) D_T.

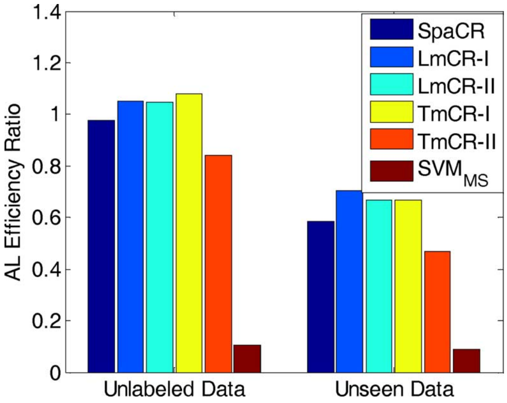


Fig. 5. AL Efficiency Ratio (*ER*) for D_U and D_T of BOT data (definitions in Table III).

TABLE IV
AL Efficiency Ratio (ER) and Incremental Change in Classification Accuracy (D) Relative to Random Sampling of SpaCR, LmCR and TmCR for BOT Data Under Different Manifold Settings

BOT	Unlabeled Data ($\mathrm{D_U}$)		Unseen Data ($\mathrm{D_T}$)	
	ER	*D*	*ER*	*D*
SpaCR	0.98	2.53	0.59	1.10
$\mathrm{LmCR_I}$	1.05	2.63	0.71	1.24
$\mathrm{LmCR_{II}}$	1.05	2.63	0.67	1.20
$\mathrm{TmCR_I}$	1.08	2.67	0.67	1.20
$\mathrm{TmCR_{II}}$	0.84	2.35	0.47	0.95

reduce the online computational complexity as compared to SpeCR during the iterative learning.

Table V also shows that in most cases, LmCR and TmCR have better performance than the spatial-based method SpaCR.

TABLE V
Incremental Change in Classification Accuracy (D) Relative to Random Sampling for KSC Data Under Different Manifold Settings

Unlabeled Data ($\mathrm{D_U}$)										
	***d*=8**			***d*=15**			***d*=2**	***d*=*B***	**AMD**	**$\mathrm{SVM_{MS}}$**
ρ	**10**	**20**	**30**	**10**	**20**	**30**	**SpaCR**	**SpeCR**		
k	**LmCR**									
5	2.86	3.11	3.22	3.34	2.99	3.59	2.87	2.91	2.27	1.85
10	2.61	3.12	3.39	3.70	3.17	3.65	2.93	3.50		
15		3.08	3.11		3.23	3.64	2.84	3.19		
20		3.18	3.53		2.79	3.68	2.88	2.84		
Mean	**2.73**	**3.12**	**3.31**	**3.52**	**3.04**	**3.64**	**2.88**	**3.11**	**2.27**	**1.85**
	TmCR									
5	2.71	3.22	3.36		3.39	3.20	2.87	2.91	2.27	1.85
10	2.70	3.04	3.24		2.94	3.32	2.93	3.50		
15		3.47	3.25		3.15	**3.96**	2.84	3.19		
20		2.98	3.29		3.12	3.75	2.88	2.84		
Mean	**2.70**	**3.18**	**3.28**		**3.15**	**3.57**	**2.88**	**3.11**	**2.27**	**1.85**

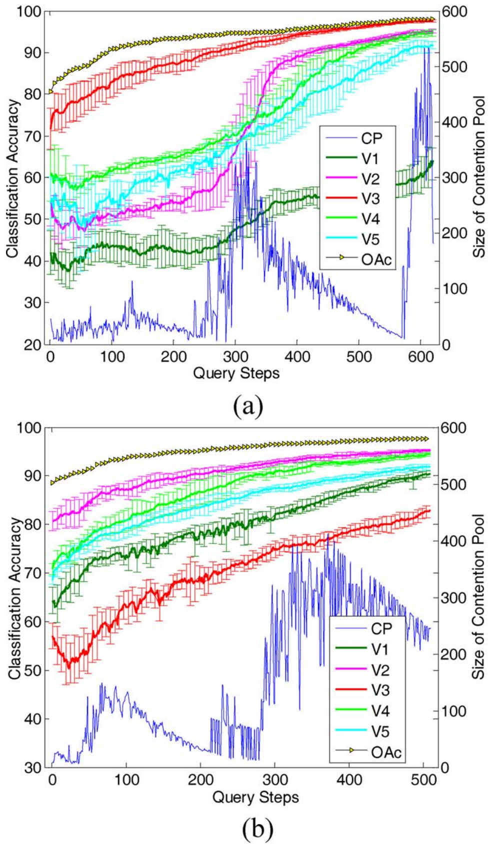
Unseen Data ($\mathrm{D_T}$)										
	***d*=8**			***d*=15**			***d*=2**	***d*=*B***	**AMD**	**$\mathrm{SVM_{MS}}$**
ρ	**10**	**20**	**30**	**10**	**20**	**30**	**SpaCR**	**SpeCR**		
k	**LmCR**									
5	1.37	1.72	1.55	2.03	1.53	2.04	1.43	1.30	1.32	1.28
10	1.01	1.60	1.71	2.07	1.55	2.02	1.51	1.68		
15		1.65	1.45		1.68	1.79	1.44	1.58		
20		1.77	1.71		1.34	1.97	1.48	1.10		
Mean	**1.19**	**1.69**	**1.60**	**2.05**	**1.52**	**1.95**	**1.47**	**1.42**	**1.32**	**1.28**
	TmCR									
5	1.29	1.87	1.56		1.61	1.80	1.43	1.30	1.32	1.28
10	1.26	1.79	1.55		1.39	1.50	1.51	1.68		
15		2.01	1.53		1.70	**2.27**	1.44	1.58		
20		1.61	1.70		1.48	2.02	1.48	1.10		
Mean	**1.28**	**1.81**	**1.58**		**1.54**	**1.90**	**1.47**	**1.42**	**1.32**	**1.28**

This is not surprising since the spatial manifold assumption is much weaker and can only be utilized effectively where good spatial clustering exists in the data, whereas the spectral-based manifold space is generated from the spectral features which have greater discriminative potential for the classification task. Furthermore, since only the spectral features are used as inputs to the classifier, LmCR and TmCR are able to adjust the conditional probability of the classifier more directly.

SpaCR generally performs slightly worse on D_U compared to SpeCR, but has comparable performance on D_T. Figs. 2 and 3 show that both KSC and BOT data have good spatial class clusters, which indicates that the spatially closeness measurement can be a good distance metric to capture the "locality" and used to identify the "close" samples. Considering that SpaCR only uses $d = 2$ features (compared with $B = 176$ in SpeCR for KSC), these results indicate that spatially based regularization might be useful for such data sets.

3) Performance of Individual Views: Fig. 6 shows an example of the classification accuracy of each view by TmCR for D_U of KSC and BOT data, respectively, where KSC experiments continued to 620 epochs in order to illustrate the final convergence of different views. The yellow curve denotes the overall accuracy obtained by the base learner, and the dark blue curve (CP) represents the size of the contention pool by the first stage regularizer. As learning progresses, the accuracy of each view is successfully bootstrapped, and different views tend to agree with each other, indicating that the degree of confusion of the learner committee on the unlabeled data decreases. The agreement of hypotheses generated from different views represents the v-intersection of those hypothesis spaces (version space) [10], [13]. The greater the agreement, the smaller the version space and closer the learner approaches the target function. The upper bound of the generalization error which is proportional to the complexity of this version space can therefore be reduced as well [12]. The three jumps of the size of the contention pool in Fig. 6(a) for KSC data around the 130th, the 320th, and the 610th queries correspond to the decrease of the maximum disagreement level from 5 to 4, 4 to 3, and 3 to 2, respectively. Similar results are also obtained for BOT data and are illustrated in Fig. 6(b).

Fig. 6 also shows that the views exhibit different discriminative ability, and performance for the two data sets differs. For KSC, View 3 (V3) has the best performance over all views, whereas View 1 has the worst performance. Table VI lists the spectral characteristics of each view for the KSC data. Views 3, 4, and 5 lie in the Red, or (N/SW) IR spectral range, while View 2 includes bands from the Green and Red part of the spectrum. View 1 contains bands in the Blue region of spectrum. Because the KSC data are dominated by green vegetation, spectral bands in the Red and NIR region have good discriminative ability, explaining the good performance of Views 2 and 3. Views 4 and 5 provide additional discrimination between upland tree classes. BOT land cover is not as dense as for KSC, and savanna grasslands and mopane woodlands have distinctive features in the SWIR portion of the spectrum. View 1 includes a continuum of bands from the Blue, Green, and Red wavelengths, while Views 2 to 5 contain groups of sequential bands from the Red, NIR, and SWIR part of the spectrum. The trend in view performance is similar, although views 2, 4, and 5 produce superior results to Views 1 and 3.



(a)

(b)

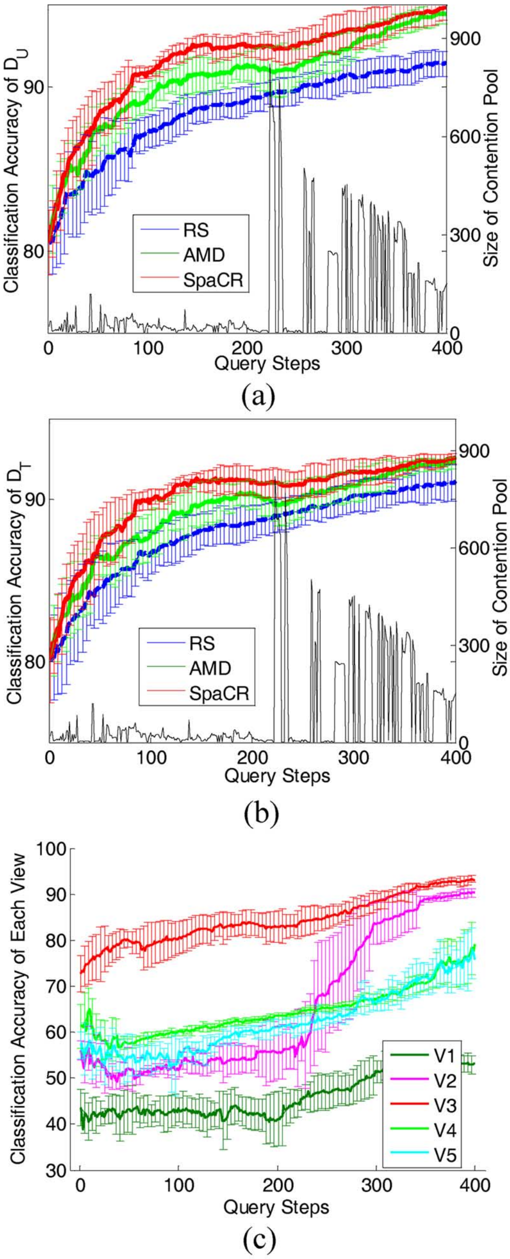
Fig. 6. Learning performance by TmCR of each view (V1-V5) and the overall classification accuracy (OAc) on D_U for (a) KSC. (b) BOT, compared with the size of the contention pool (CP).

TABLE VI
SPECTRAL CHARACTERISTICS OF EACH VIEW FOR KSC DATA

View	Band Indices	Wavelength Range (nm)	Spectral Range
1	1 - 11	409.21 - 507.74	Blue
2	12 - 31	517.60 - 683.30	Green + Red
3	32 - 96	692.88 - 1284.20	Red + NIR
4	97 - 130	1294.15 – 1772.06	NIR + SWIR
5	131 - 176	1782.02 – 2437.40	SWIR

Blue: 435-500nm; **Green**: 520-565nm; **Red**: 625-750nm
Near Infrared (NIR): 750nm-1500nm;
Short-wave IR (SWIR) : 1500nm-3000nm

4) Contribution of the Multi-View Regularization Versus the Local Proximity Regularization: Fig. 7(a) and (b) shows the classification accuracies of AMD, SpaCR, and RS for D_U and D_T of KSC data, respectively. The black curve at the bottom in both figures denotes the size of the contention pool from one test case. Fig. 7(c) shows the corresponding classification accuracy of each view by AMD. Since the two stage regularizers perform sequentially, while the performance is improved by the second stage regularizer, it is also affected by the first stage regularizer. This phenomenon can be seen clearly at the $\sim$ 220th step of querying, where both SpaCR and AMD exhibit a slower improvement as learning progresses. This is due to the decrease of the maximum disagreement level as View 2 has improved as shown in Fig. 7(c), which leads to inflation of the contention pool. Since more samples with lower disagreement level are incorporated into the contention pool, by randomly sampling from this contention pool, the



(a)

(b)

(c)

Fig. 7. Comparison of AMD-SR, SpaCR, and RS for KSC data, with the size of the corresponding contention pool (black curves): performance on (a) D_U, and (b) D_T, (c) classification accuracy of each view by AMD-SR on D_U.

learner may not be able to focus on the most informative samples, thus resulting in slower improvement of the classification accuracy. However, with additional local consistency evaluation, samples in this pool are further ranked, which helps SpaCR to focus on the more informative samples. Thus, the degradation in the performance is less for SpaCR than for AMD.

D. Computational Complexity

The overall computational complexity of the method consists of two parts. The offline part consists of computing the manifold space and finding the neighboring samples in the new manifold space, which mainly depends on the complexity of the manifold learning method. Both LLE and LTSA scale as $O(BN^2 \log(N))$, where $N = N_L + N_U$ is the total sample number of the labeled and the unlabeled data. Further, the search of the k-nearest neighbor samples in the new generated manifold space scales as $O(dN^2 \log(N))$.

The online active learning component of the method depends on the total number queried N_q, the efficiency of the base learner, the size of the k-nearest neighborhood for computing the local proximity, the dimension of the generated manifold space, and the size of the unlabeled data pool.

In our case, at the τth query, the base learner SVM requires $O\left(N_v (N_L^\tau)^2\right) \sim O\left(N_v (N_L^\tau)^3\right)$ for training, which depends on the number of support vectors $N_{sv}^\tau (N_{sv}^\tau \leq N_L^\tau)$, and $O(N_v N_U^\tau N_{sv}^\tau)$ for predicting on the unlabeled data.

Since we apply the adaptive maximum disagreement regularizer $\mathfrak{R}_{\mathrm{AMD}}$ to build the first stage contention pool, the number of candidate samples is greatly reduced, and it is only necessary to evaluate those candidate samples by the second stage regularizer, thus reducing the computational load. Denote N_{CP} as the number of candidate samples in C_{AMD}, where usually $N_{\mathrm{CP}} \ll N_U$. The calculation of the regularizer $\mathfrak{R}_{\mathrm{LIC}}$ for each candidate sample and sorting the results to identify the next query sample scales as $O(kdN_{\mathrm{CP}} \log(N_{\mathrm{CP}}))$, which further depends on the dimension of the feature used for computing the distance between $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ in (13). By nonlinear mapping of the original high dimensional data into the low-dimensional manifold space, a more reliable distance measure is obtained, and also since $d \ll B$ the online computational complexity is improved.

Since our method does not require a specific learner, additional computational improvement can be achieved by using a more efficient base learner. This also provides the flexibility to choose the proper model for different types of data.

VI. Conclusion

In this paper, we have presented a new sequential co-regularization active learning framework that utilizes multi-view consistency and the local proximity assumption for remote sensing image classification. The first regularizer explores the intrinsic multi-view information embedded in the hyperspectral image to boost learning by using the complementary information from the disjoint feature subspaces. The second regularizer seeks boundary samples in the spatial/spectral low dimensional manifold structure, as those samples are more useful for improving the discriminative SVM classifier. Based on different manifold assumptions, three methods are developed which emphasize local consistency: SpaCR, LmCR, and TmCR. Experiments show the improvement achieved by adding the second stage regularizer to the AMD regularizer. As compared to SpeCR, which is based on the original high-dimensional spectral feature space, manifold based methods achieve higher classification accuracies and greater efficiency, with less online computational cost due to the resulting low-dimensional manifold space. All the methods, including AMD and SpeCR, show excellent performance as compared to RS and $\mathrm{SVM}_{\mathrm{MS}}$ on two data sets (KSC and BOT). This indicates the potential effectiveness of the regularization based active learning framework. The good performance by AMD, where views are obtained according to correlation between spectral bands, motives us to further investigate the properties associated with multi-view learning and alternative ways of generating views. Our results also indicate that class dependent manifold embedding may be a promising direction for future research.

References

[1] M. Ferecatu and N. Boujemaa, "Interactive remote-sensing image retrieval using active relevance feedback," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 818–826, Apr. 2007.

[2] P. Mitra, B. U. Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognition Lett.*, vol. 25, no. 9, pp. 1067–1074, Jul. 2004.

[3] 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.

[4] 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.

[5] A. Liu, G. Jun, and J. Ghosh, "A self-training approach to cost sensitive uncertainty sampling,," *Mach. Learn. J.*, vol. 76, no. 2, pp. 257–270, 2009.

[6] W. Di and M. Crawford, "Locally consistent graph regularization based active learning for hyperspectral image classification," in *Proc. 2nd IEEE Workshop Hyperspectral Image Signal Process.: Evolut. Remote Sens.*, Reykjavik, Iceland, 2010, pp. 1–4.

[7] W. Di and M. Crawford, "Multi-view adaptive disagreement based active learning for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Honolulu, HI, 2010, pp. 1374–1377.

[8] G. Schohn and D. Cohn, "Less is more: Active learning with support vectors machines," in *Proc. 17th Int. Conf. Mach. Learn.*, Stanford, CA, Jul. 2000, pp. 839–846.

[9] B. Settles, "Active learning literature survey," Wisconsin-Madison, 2009, Computer Sciences Tech. Rep. 1648.

[10] I. Muslea, S. Minton, and C. A. Knoblock, "Active learning with multiple views," *J. Artif. Intell. Res.*, vol. 27, pp. 203–233, 2006.

[11] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res. Archive*, vol. 2, pp. 45–66, Mar. 2002.

[12] S. Abney, "Bootstrapping," in *Proc. 40th Meeting Assoc. Comput. Linguist.*, Philadelphia, PA, Jul. 2002, pp. 360–367.

[13] B. Leskes, "The value of agreement a new boosting algorithm," M.S. thesis, Univ. of Amsterdam, Amsterdam, The Netherlands, 2002.

[14] F. Olsson, "A literature survey of active machine learning in the context of natural language processing," Swedish Inst. of Comput. Sci., SICS Tech. Rep., 2009.

[15] V. R. de Sa, "Learning classification with unlabeled data," *Adv. Neural Inf. Process. Syst.*, vol. 6, pp. 112–119, 1994.

[16] Z. Wang and S. Chen, "Multi-view kernel machine on single view data," *Neurocomputing*, vol. 72, no. 10–12, pp. 2444–2449, Jun. 2009.

[17] I. Muslea, S. Minton, and C. Knoblock, "Active + semi-supervised learning = robust multi-view learning," in *Proc. 19th Int. Conf. Mach. Learn.*, Sydney, Australia, Jul. 2002, pp. 435–442.

[18] L. Gómez-Chova, G. Camps-Valls, L. Bruzzone, and J. Calpe-Maravilla, "Mean map kernel methods for semisupervised cloud classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 207–220, Jan. 2010.

[19] J. A. Lee and M. Verleysen*, Nonlinear Dimensionality Reduction*. New York: Springer, 2007, Information science and statistics.

[20] O. Bousquet, O. Chapelle, and M. Hein, "Measure based regularization," in *Adv. in Neural Inf. Process. Syst.*. Cambridge, MA: MIT Press, 2004, vol. 16.

[21] X. Zhu, "Semi-supervised learning literature survey," Univ. of Wisconsin-Madison, Comput. Sci., 2005, Tech. Rep. 1530.

[22] C. M. Bachmann and T. L. Ainsworth, "Exploiting manifold geometry in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 441–454, Mar. 2005.

[23] T. Han and D. G. Goodenough, "Nonlinear feature extraction of hyperspectral data based on locally linear embedding," in *IEEE Int. Conf. Symp. Geosci. Remote Sens. Symp.*, Seoul, Korea, Jul. 2005, vol. 2, pp. 1237–1240.

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

[25] 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, Dec. 2004.

[26] W. Kim and M. M. Crawford, "Adaptive classification for hyperspectral image data using manifold regularization kernel machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4110–4121, Nov. 2010.

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

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

[29] X. Jia and J. A. Richards, "Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 1, pp. 538–542, Jan. 1999.

[30] A. Hinneburg, C. C. Aggarwal, and D. A. Keim, "What is the nearest neighbor in high dimensional spaces?," in *Proc. 26th Int. Conf. Very Large Data Bases*, Cairo, Egypt, Sep. 2000, pp. 506–515.

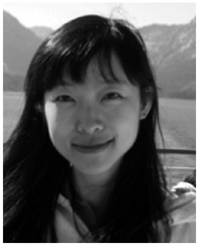
[31] K. Beyer, J. Goldstein, R. Ramakishnan, and U. Shaft, "When is nearest neighbor meaningful?," in *Proc. 7th Int. Conf. Database Theory*, Jerusalem, Israel, Jan. 1999, pp. 217–235.

[32] A. L. Neuenschwander, "Remote sensing of vegetation dynamics in response to flooding and fire in the Okavango Delta, Botswana," Ph.D. dissertation, Univ. of Texas at Austin, Austin, TX, 2007.

[33] J. Ham, Y. Chen, M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[34] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," 2001 [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[35] G. Camps-Valls, L. Gómez-Chova, J. Calpe, E. Soria, J. D. Martín, L. Alonso, and J. Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1530–1542, Jul. 2004.



Wei Di (M'07) received the M.S. degree in control theory and engineering from Northwestern Polytechnical University, Xi'an, China, in 2008. She is currently pursuing the Ph.D. degree in civil engineering, geomatics, and also enrolled into the interdisciplinary program of *Computational Science and Engineering (CS&E)* at Purdue University, West Lafayette, IN.

Her research interests include data mining, active learning, image classification, statistical pattern recognition, information retrieval, and related applications for hyperspectral data analysis.

Melba M. Crawford (M'90–SM'05–F'08) received the B.S. and M.S. degrees in civil engineering from the University of Illinois, Urbana, and the Ph.D. degree in systems engineering from Ohio State University, Columbus.

She was a faculty member at the University of Texas at Austin from 1990 to 2005, where she founded an interdisciplinary research and applications development program in space-based and airborne remote sensing. She is currently at Purdue University, West Lafayette, IN, where she holds the Purdue Chair of Excellence in Earth Observation, is Director of the Laboratory for Applications of Remote Sensing, and Associate Dean of Engineering for Research. Her research interests focus on hyperspectral and lidar sensing, data fusion for multisensor problems, manifold learning, and knowledge transfer in data mining.

Dr. Crawford was a Jefferson Senior Science Fellow at the U.S. Department of State from 2004 to 2005 and continues to serve in an advisory capacity. She is Executive Vice President of the IEEE Geoscience and Remote Sensing Society. She also served as a member of the NASA Earth System Science and Applications Advisory Committee (ESSAAC) and was a member of the NASA EO-1 Science Validation team for the Advanced Land Imager and Hyperion. She is currently a member of an advisory committee for the IEEE Committee on Earth Observation to the South African Department of Science and Technology for capacity building in space technologies and remote sensing applications.