An Active Learning Approach to Hyperspectral Data Classification

Suju Rajan, Joydeep Ghosh, *Fellow, IEEE*, and Melba M. Crawford, *Fellow, IEEE*

***Abstract*—Obtaining training data for land cover classification using remotely sensed data is time consuming and expensive especially for relatively inaccessible locations. Therefore, designing classifiers that use as few labeled data points as possible is highly desirable. Existing approaches typically make use of small-sample techniques and semisupervision to deal with the lack of labeled data. In this paper, we propose an active learning technique that efficiently updates existing classifiers by using fewer labeled data points than semisupervised methods. Further, unlike semisupervised methods, our proposed technique is well suited for learning or adapting classifiers when there is substantial change in the spectral signatures between labeled and unlabeled data. Thus, our active learning approach is also useful for classifying a series of spatially/temporally related images, wherein the spectral signatures vary across the images. Our interleaved semisupervised active learning method was tested on both single and spatially/temporally related hyperspectral data sets. We present empirical results that establish the superior performance of our proposed approach versus other active learning and semisupervised methods.**

***Index Terms*—Active learning, hierarchical classifier, multitemporal data, semisupervised classifiers, spatially separate data.**

I. Introduction

RECENT advances in remote sensing technology have made hyperspectral data with hundreds of narrow contiguous bands more widely available. The hyperspectral data can therefore reveal subtle differences in the spectral signatures of land cover classes that appear similar when viewed by multispectral sensors [1]. If successfully exploited, the hyperspectral data can yield higher classification accuracies and more detailed class taxonomies. However, the task of classifying hyperspectral data also has unique challenges.

Supervised statistical methods require labeled training data to estimate parameters. It is expensive and time consuming to obtain labeled data, but the very high dimensionality of the hyperspectral data makes it difficult to design classifiers using only a few labeled data points.

The task of classifying hyperspectral images obtained over different geographic locations or multiple times proportionately becomes more complex as factors such as atmospheric and light conditions, topographic variations, etc., alter the spectral signatures corresponding to the same land cover type across different images. Rather than acquiring labeled data from each of the spatially/temporally related images, it would be very desirable to acquire labeled data from a single image and exploit that knowledge for constructing a new classifier for a new but related image. We refer to this concept of exploiting labeled data from related images as the *knowledge transfer* scenario [2].

The focus of this paper is on hyperspectral image classification using very few labeled data points. Two popular machine learning approaches for dealing with this problem are semisupervised learning and active learning. Semisupervised algorithms incorporate the unlabeled data into the classifier training phase to obtain better decision boundaries. Some of the more popular semisupervised classification algorithms are techniques based on Expectation Maximization (EM) [3] and transductive support vector machines [4]. An overview of the semisupervised classification techniques can be found in [5]. In contrast, active learning [6] assumes the existence of a rudimentary learner trained with a small amount of labeled data. The learner has access to both the unlabeled data and a "teacher." The learner then selects an unlabeled data point and obtains its label from the teacher. The goal of the active learner is to select the most "informative" data points so as to accurately learn from the fewest such additionally labeled data points. Several active learning algorithms have been proposed, which differ in the way the unlabeled data points are chosen.

In this paper, we explore the efficacy of combining semisupervision with a new active learning technique in building hyperspectral classifiers using very little labeled data. Our technique is applicable to both tasks of single-image classification and knowledge transfer. For single-image classification, we assume that we have very little labeled data from the image. We then use active learning to select unlabeled data points from the same image for retraining the classifier. In the knowledge transfer scenario, we assume that we have several temporally or spatially related images. The labeled data from one such image are used to build the initial classifier. The unlabeled data points are then selected from the temporally or spatially separate image to efficiently update the existing classifier for this separate image.

Our proposed method works with any generative classifier. In this paper, we evaluate our technique using two such classifiers, namely the Maximum-Likelihood (ML) classifier and the Binary Hierarchical Classifier (BHC). We present results on several isolated and spatially/temporally related hyperspectral

Manuscript received November 9, 2006; revised April 9, 2007. This work was supported by the National Science Foundation under Grant IIS-0312471.

S. Rajan and J. Ghosh are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: rsuju@lans.ece.utexas.edu; ghosh@lans.ece.utexas.edu).

M. M. Crawford is with the Schools of Civil and Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: 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/TGRS.2007.910220

images. In all cases, our method of incorporating semisupervision with active learning is found to perform better than other active learning approaches (also interleaved with semisupervision) and classical semisupervised methods.

II. Related Work

The classification of hyperspectral data using labeled data, including specialized techniques for dealing with small sample sizes [7], [8], has been well studied in the remote sensing community, so we do not review this literature but refer to the special issue [9]. Rather, since the focus of this paper is on learning classifiers when only a portion of the data is labeled, we focus on the existing literature on semisupervised learning and knowledge transfer for remotely sensed data. To our knowledge, there has been very little work in using active learning for the classification of remotely sensed data. Hence, our review of active learning concentrates on the general theoretical frameworks developed in the machine learning community.

A. Semisupervised Learning for Single-Image Classification

Given a mixture of labeled and unlabeled data, semisupervised classification algorithms [5] try to improve the classification accuracy by making use of the unlabeled data to obtain better classification boundaries. Semisupervised methods that make use of EM have had considerable success in a number of domains, especially that of text data analysis and remote sensing.

The advantages of using unlabeled data to aid the classification process in the domain of remote sensing data were first identified and exploited in [7]. In this paper, the authors made use of unlabeled data via EM to obtain better estimates of the class-specific parameters. It was shown that using unlabeled data enhanced the performance of the maximum *a posteriori* probability classifiers, especially when the dimensionality of the data approached the number of training samples. Subsequent extensions to the EM approach include using "semilabeled" data in the EM iterations [10], [11]. In these methods, the available labeled data are first used to train a supervised classifier to obtain tentative labels for the unlabeled data. The semilabeled data thus obtained are then used to retrain the existing classifier, and the process is iterated until convergence.

In addition to the typical semisupervised setting, unlabeled data have also been utilized for "partially supervised classification" [12], [13]. In partially supervised classification problems, training samples are provided only for a specific class of interest, and the classifier must determine whether the unlabeled data belong to the class of interest. While Mantero *et al.* [13] attempt to model the distribution of the class of interest and automatically determine a suitable "acceptance probability," Jeon and Landgrebe [12] make use of the unlabeled data while learning an ML classifier to determine whether a data point is of interest or not.

B. Semisupervised Learning for Multi-Image Classification

The possibility that the class label of a pixel could change with time was first explored in [14]. In this paper, the joint probabilities of all the possible combinations of classes between the multitemporal images were estimated and used in the classification rule. However, the proposed "multitemporal cascade classifier" requires labeled data from all the images of interest. More recently, unsupervised algorithms have been proposed whereby changes in the label of a particular pixel in a multitemporal sequence are automatically detected [15]. Supervised methods that automatically try to model the class transitions in multitemporal images have also been investigated [16]. Another supervised approach involves building a local classifier for each image in the sequence and combining the decisions, either via a joint-likelihood-based rule or a weighted majority decision rule based on the reliabilities of the data sets and that of the individual classes, to yield a "global" decision rule for the unlabeled data [17]. Still other spatial–temporal methods utilize the temporal correlation of the classes between images to help improve the classification accuracy [18], [19].

It is important to note that the standard formulation for semisupervised classification techniques assumes that both labeled and unlabeled data have the same class-conditional distributions. This assumption is violated for the knowledge transfer scenario considered in this paper. While applying a classifier learned on a particular image to a spatially/temporally separated image, it is likely that the statistics of the data from the new images significantly differ from the original image. The standard semisupervised approach may not be the best option.

C. Knowledge Transfer for Classification of Related Images

A pioneering attempt at unsupervised knowledge transfer for multitemporal remote sensing images was made in [20]. In this paper, the authors consider a fixed set of land cover classes whose spectral signatures vary over time. Given an image t_1 of a certain land area with a labeled training set, the problem is to classify pixels of another image t_2 of the same land area obtained at a different time. An ML classifier is first trained on the labeled data from t_1 assuming that the class-conditional density functions are Gaussian. The mean vector and the covariance matrix of the classes from t_1 are used as initial approximations to the parameter values of the same classes from t_2. These initial estimates to the classes from t_2 are then improved via EM using the corresponding unlabeled data.

A recent work exploited the "contextual" properties of a classifier trained using data acquired from one area to help classify the data obtained from spatially and temporally different areas [2]. A multiclassifier system called the BHC [21] was used for this purpose. The BHC automatically derives a hierarchy of the target classes based on their mutual affinities. This hierarchy, along with the features extracted at each node of the BHC tree, is used to transfer the knowledge from an existing classification task to another related task. The available unlabeled data are then used to update the existing BHC via semisupervised learning techniques to better reflect the statistics of the data from new areas. It was shown that exploiting contextual information yielded better classification accuracies than other powerful multiclassifier systems, such as the error-correcting output code [22], for the purposes of knowledge transfer in hyperspectral data.

D. Active Learning

In a typical active learning setting, a classifier is first trained from a small amount of labeled data. The classifier also has access to the set of unlabeled data as well as a "teacher." The classifier then selects a data point from the set of unlabeled data points and obtains the corresponding label from the teacher. The goal of the algorithm is to choose data points such that a more accurate classification boundary is learned using as few additional labeled data points as possible. Stated formally, let $\mathcal{X} \in \Re^{(n\times m)}$ be a random vector following a certain probability distribution, for example, P_X. Assume that the learner has access to a set of random instances $D = \{x_i\}_{i=1}^{n}$ drawn from P_X. Let $D_L \subset D$ be the subset for which the true target value $\{y_i\}_{i=1}^{n}$ has been provided to train a classifier. Active learning algorithms then select $\hat{x}$ from $D_{UL} = D \backslash D_L$ and retrain the classifier with the appended training set $D_L^+ = D_L \cup (\hat{x}, \hat{y})$. Note that the learner does not have access to the label $\hat{y}$ prior to committing to a specific $\hat{x}$. The process of identifying $\hat{x}$ and adding it to D_L is repeated for a user-specified number of iterations. The different active learning methods differ in the criteria used to select $\hat{x}$.

The only work that applies active learning for the classification of remotely sensed data that we are aware of at this time is that of Mitra *et al.* [23], which restricts itself to multispectral images. In [23], Mitra *et al.* make use of active learning while training support vector machines, and identify $\hat{x}$ from D_{UL} based on the distance of the unlabeled data points from the existing hyperplane. The dependence of the selection criterion for $\hat{x}$ on the hyperplane limits the approach to Support Vector Machine (SVM) type classifiers.

A statistical approach to active learning for regression problems was proposed by Cohn *et al.* [24], where bias-variance decomposition of the squared error function is used to select $\hat{x}$. Assuming an unbiased learner, $\hat{x}$ is selected such that the resulting D_L^+ minimizes the expected variance in the output of the learner measured over X, where the expectation is taken over $P(Y|\hat{x})$.

In a related work, MacKay [6] proposed an information-based objective function for active learning. In this setting, the true target function is characterized by a parameter vector w over which a probability distribution $P(W|D)$ is defined. Defining S as the entropy of $P(W|D_L)$ and S^+ as the entropy with $P(W|D_L^+)$, the goal is to select $\hat{x}$ such that the expected change in entropy of the distribution $(S - S^+)$ is maximum. The authors also show that maximizing the expected change in entropy is the same as maximizing the Kullback–Leibler (KL) divergence between $P(W|D_L^+)$ and $P(W|D_L)$ [25]. Under the regression setting, it is shown that choosing $\hat{x}$ as the point for which the estimated target value (based on w) has the maximum variance causes the maximum increase in mutual information of the parameter vector. The authors also present a closed-form solution in the regression setting for identifying $\hat{x}$ with the maximum information gain.

An active learning approach that makes use of the *a posteriori* probability density function (pdf) $(P(Y|X))$ was proposed in [26]. Assuming there exists a true probability distribution $(P_{\text{true}}(Y|X))$, a user-defined loss function $\mathcal{L}$, and the *a posteriori* probability distribution estimated from the training set $(P_{D_L}(Y|X))$, the expected loss of the learner is defined as

$$E_{D_L}(L) = \int_x \mathcal{L}\left(P_{\text{true}}(Y|x), P_{D_L}(Y|x)\right) P(x)dx. \quad (1)$$

Active learning proceeds by selecting a data point such that the expected error using the appended training set D_L^+ is the least over all the possible $\hat{x} \in D_{UL}$. However, since $P_{\text{true}}(Y|X)$ is unknown, the authors propose using $P_{D_L}(Y|X)$ itself as an estimate for the unknown true distribution. This substitution renders the expected loss function meaningless when L is chosen to be the Euclidean distance. When using the KL divergence as the loss function, the equation reduces to the negative entropy of $P_{D_L}(Y|X)$ [26].

Given a probabilistic binary classifier, the uncertainty sampling technique proposed by Lewis and Gale [27] chooses the data point whose *a posteriori* estimates $P_{D_L}(y|\hat{x})$ are closest to 0.5. Since the method focuses on examples closer to the decision boundaries, it is not clear whether this method will be of much use for data sets with considerable overlap between classes as data points close to the decision boundary will always be chosen for labeling, which results in skewed class-conditional probability estimates.

Committee-based learners comprise another popular class of "multihypothesis" active learning algorithms. Of these methods, the "query by committee" (QBC) approach in [28] is a general active learning algorithm that has theoretical guarantees on the reduction in prediction error with the number of queries. Given an infinite stream of unlabeled examples, the QBC picks the data point on which instances of the Gibbs algorithm, which are drawn according to a probability distribution defined over the version space, disagree. However, the algorithm assumes the existence of a Gibbs algorithm and noise-free data. Several variations of the original QBC algorithm have been proposed, such as the Query by Bagging and Query by Boosting algorithms [29] and the adaptive resampling approach [30].

Saar-Tsechansky and Provost apply active learning principles to obtain better class (*a posteriori*) probability estimates [31]. Given a probabilistic classifier, the Bootstrap-LV attempts to select $\hat{x}$ with the highest "local variance" assuming that the example that has a high variance in its class probability estimate is more difficult to learn and hence should be queried. An extension of the Bootstrap-LV algorithm to the multiclass case [32] makes use of the Jensen–Shannon divergence to measure the uncertainty in the class probability estimates.

McCallum and Nigam [33] combine EM and active learning for text classification. Based on the QBC approach, the "density-weighted pool-based sampling" uses the average KL divergence between the *a posteriori* class distribution of each classifier and the mean *a posteriori* class distribution (i.e., the Jensen–Shannon divergence with equal weights) to assign a disagreement score to each x in D_{UL}. The disagreement measure is then combined with a density metric that ensures that the algorithm chooses an $\hat{x}$ that is similar to many other data points in D. Thus, each $\hat{x}$ is not only representative of other

data points but also causes significant disagreement among the committee members.

Active learning has also been applied in the multiview setting [34]. In the multiview problem, the features are partitioned into subsets, each of which is sufficient for learning an estimate of the target function. In the co-testing family of algorithms, classifiers are constructed for each view of the data. Provided the views are "compatible" and "uncorrelated," the data points on which the classifiers disagree are likely to be most informative.

III. Proposed Approach

We propose a new active learning technique that can be used in conjunction with any classifier that determines the decision boundary via (an estimate of) *a posteriori* class probabilities, i.e., classifiers that are probabilistic/generative rather than discriminative [35].

Our approach strikes a middle ground between the methods proposed in [6] and that in [26]. As in [26], we make use of the *a posteriori* probability distribution function $P(Y|X)$ to guide our active learning process. The loss function we propose is similar to that in [6] in that we attempt to increase the information gain between $P_{D_L^+}(Y|X)$ and $P_{D_L}(Y|X)$, i.e., the *a posteriori* pdfs estimated from D_L^+ and D_L, respectively. Maximizing the expected information gain between $P_{D_L}^+(Y|X)$ and $P_{D_L}(Y|X)$ is equivalent to selecting the data point $\hat{x}$ from D_{UL} such that the expected KL divergence between $P_{D_L}^+(Y|X)$ and $P_{D_L}(Y|X)$ is maximized. That is, we try to select those data points that change the current belief in the posterior probability distribution the most.

Since the true label of $\hat{x}$ is initially unknown, we follow the methodology in [24] and [26] and estimate the expected KL distance between $P_{D_L}^+(Y|X)$ and $P_{D_L}(Y|X)$ by first selecting $\tilde{x} \in D_{UL}$ and assuming $\tilde{y}$ to be its label. Let $D_{UL}^+ = D_{UL} \backslash \tilde{x}$, $D_L^+ = D_L \cup (\tilde{x}, \tilde{y})$, and $|D_{UL}^+|$ be the number of data points in the set D_{UL}^+. Estimating via sampling, the proposed $KL^{\max}$ function can be written in terms of $(\tilde{x}, \tilde{y})$ as

$$KL_{D_L^+}^{\max}(\tilde{x}, \tilde{y}) = \frac{1}{|D_{UL}^+|} \sum_{x \in D_{UL}^+} KL\left(P_{D_L}^+(Y|x) \| P_{D_L}(Y|x)\right). \tag{2}$$

The KL divergence between the two probability distributions is defined as

$$KL\left(P_{D_L}^+(Y|x) \| P_{D_L}(Y|x)\right) = \sum_{x \in D_{UL}^+} P_{D_L}^+(Y|x) \log\left(\frac{P_{D_L}^+(Y|x)}{P_{D_L}(Y|x)}\right). \tag{3}$$

Note that simply assigning a wrong class label to $\tilde{y}$ for $\tilde{x}$ can result in a large value of the corresponding $KL_{D_L^+}^{\max}$. Hence, as in [24] and [26], we use the expected KL distance from $P_{D_L}^+(Y|x)$ and $P_{D_L}(Y|x)$, with the expectation estimated over $P_{D_L}(Y|x)$, and then select the $\hat{x}$ that maximizes this distance as

$$\hat{x} = \underset{\tilde{x} \in D_{UL}}{\operatorname{argmax}} \sum_{\tilde{y} \in Y} KL_{D_L^+}^{\max}(\tilde{x}, \tilde{y}) P_{D_L}(\tilde{y}|\tilde{x}). \tag{4}$$

The efficacy of our method strongly depends on the correctness of the posterior probability estimates. The very high dimensionality (>100 features) of the hyperspectral data coupled with the lack of sufficient quantities of labeled data could result in skewed estimates of the parameters of the probability distributions. The dimensionality of the data is reduced via feature selection/extraction techniques [36], and the EM algorithm is utilized with the active learning process to improve the estimates.

The following subsections describe our method in more detail for the two different application scenarios, i.e., classifying a single hyperspectral image and knowledge transfer between multiple temporally/spatially related images. We use an ML classifier and our own BHC, in which each class is modeled by a multivariate Gaussian. However, it should be clear that our technique can be used with any classifier that can produce estimates of *a posteriori* class probabilities.

A. Active Learning for Classifying a Single Image

Let us assume that we have a small amount of labeled data from the hyperspectral image to be classified. The high-dimensional data are first projected into a reduced space using feature selection/extraction techniques. We choose the Fisher-m feature extractor [37] for the following reasons: 1) the Fisher extractor produces a feature space that is most suitable for discriminating the different land cover classes; and 2) the Fisher discriminant makes use of the estimates of the class distributions to determine the reduced space and can be continually updated to reflect the changes in the estimates as the learning proceeds.

When using the ML classifier with multivariate Gaussians to model the class-conditional probability distributions, the initial parameters of the Gaussians are estimated using the available labeled data. The E-step of the algorithm determines the posterior probabilities of the unlabeled data based on the Gaussians. The probabilities thus estimated are then used to update the parameters of the Gaussians (M-step). EM iterations are performed until the average change in the posterior probabilities between two iterations is smaller than a specified threshold [20]. A new Fisher feature extractor is computed for each EM iteration based on the statistics of the classes at that iteration. The updated extractor can then be used to project the data into the corresponding Fisher space prior to the estimation of the class-conditional pdfs.

Setting $P_{D_L}(Y|X)$ as the posterior probability of the unlabeled data D_{UL}, which is obtained at the end of the EM iterations, $(\hat{x}, \hat{y})$ is selected from D_{UL} such that the expected KL divergence between $P_{D_L}^+(Y|X)$ and $P_{D_L}(Y|X)$ is maximized, where $D_L^+ = D_{UL} \cup (\hat{x}, \hat{y})$. For reasons of computational efficiency, $(\hat{x}, \hat{y})$ is selected from a randomly sampled subset of D_{UL}. A data point $\tilde{x}$ is selected from the subset of D_{UL}, and

the label $\tilde{y}$ is assigned to it. This new data point $(\tilde{x}, \tilde{y})$ is then used to update the existing class parameter estimates, and a new posterior probability distribution $P^{+}_{D_L}(Y|X)$ is obtained. Using (2) and (4), the expected value of $KL^{\max}_{D^{+}_{L}}(\tilde{x}, \tilde{y})$ is computed over $D^{+}_{UL} = D_{UL} \backslash \tilde{x}$ for all possible $\tilde{y}$. The data point $(\hat{x}, \hat{y})$ from D_{UL} with the maximum expected KL divergence is then added to the set of labeled data points, where $\hat{y}$ is hereafter assumed to be the true label of $\hat{x}$.

For the next iteration of active learning, the EM process is repeated but with two differences: 1) the Gaussian parameter estimates from the previous iteration are used to initialize the EM process, and 2) constrained EM is employed, wherein the E-step only updates the posterior probabilities for the unlabeled data while fixing the memberships of the labeled instances according to the known class assignments.

B. Active Learning for Knowledge Transfer

Assume that the hyperspectral data are available from two spatially (or temporally) different areas, i.e., Areas 1 and 2, and that there is an adequate amount of labeled data from Area 1 to build a supervised classifier. The Fisher-m feature extractor is computed from the Area 1 data to determine a low-dimensional discriminatory feature space.

The one difference between active learning for the single-image case and that of the knowledge transfer scenario is that in the latter the unlabeled data are drawn from spatially/temporally removed data. While the labeled data from Area 1 are only used to initialize the very first EM iteration, subsequent EM iterations are guided by the posterior probabilities assigned to the unlabeled Area 2 data. Active learning proceeds as before with the posterior probability distributions of the Area 2 data determining $P_{D_L}(Y|X)$ and guiding the active learning process. Thus, we ensure that we select "informative" Area 2 data points that change the existing belief in the distributions of the Area 2 classes the most. Selecting such data points should result in better learning curves than if the data are selected at random. Constrained EM is then performed between active learning iterations by using the estimates from the previous EM iteration for initialization and holding the known memberships of the Area 2 data points as fixed.

IV. Experimental Evaluation

Results were obtained to investigate the performance of our proposed method. We compared the learning rates with those of other classifiers that select data points either at random or via another related active learning method.

A. Data Sets

The active learning approaches described above were tested on hyperspectral data sets obtained from two sites: the John F. Kennedy Space Center (KSC), National Aeronautics and Space Administration (NASA), Florida [38], and the Okavango Delta, Botswana [39]. The images of the data sets along with the spatial regions from which the labeled data were obtained are shown in [40].

TABLE I
Class Names and Number of Data Points for the KSC Data Set

No.	Class Name	Area 1	Area 2
1.	Scrub	761	422
2.	Willow Swamp	243	180
3.	CP Hammock	256	431
4.	CP/Oak Hammock	252	132
5.	Slash Pine	161	166
6.	Oak/ Broadleaf Hammock	229	274
7.	Hardwood Swamp	105	248
8.	Graminoid Marsh	431	453
9.	Salt Marsh	419	156
10.	Water	927	1392

1) KSC: The NASA Airborne Visible/Infrared Imaging Spectrometer acquired data at 18-m spatial resolution over the KSC on March 23, 1996. The bands that were noisy or impacted by water absorption were removed, which leaves 176 candidate features for the study. The training data were selected using land cover maps derived by the KSC staff from color infrared photography, Landsat Thematic Mapper (TM) imagery, and field checks. The discrimination of land cover types for this environment is difficult due to the similarity of the spectral signatures for certain vegetation types and the existence of mixed classes. The 512 $\times$ 614 spatially removed data set is located on a different portion of the flight line and exhibits somewhat different characteristics [40]. While the number of classes in the two regions differs, we restrict the study to those classes that are present in both regions. Details of the ten land cover classes considered in the KSC area are shown in Table I.

2) Botswana: Hyperion data were acquired over a 1476 $\times$ 256 pixel study area located in the Okavango Delta, Botswana. Fourteen different land cover types consisting of seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the delta were identified for the study, which focused on the impact of flooding on vegetation. Uncalibrated and noisy bands that cover water absorption features were removed, which results in 145 features. The training data were manually selected using a combination of vegetation surveys located by the Global Positioning System, aerial photography from the Aquarap (2000) project, and a 2.6-m resolution IKONOS multispectral imagery. The spatially removed test data for the May 31, 2001 acquisition were sampled from spatially contiguous clusters of pixels that were within the same scene but disjoint from those used for the training data [40]. Table II contains a list of classes and the number of class-specific labeled data.

Multitemporal data: To test the efficacy of the knowledge transfer framework for multitemporal images, data were also obtained from the Okavango region in June and July 2001. The May data are characterized by the onset of the annual flooding cycle and some newly burned areas. The flood progressed in June and July, and the burned vegetation recovered. It should also be noted that only nine classes were identified in the June and July images as the data were acquired over a slightly different area due to a change in the satellite pointing. Additionally, some classes identified in the May 2001 image were excessively fine grained for this sequence, so the data in some classes were

TABLE II
CLASS NAMES AND NUMBER OF DATA POINTS FOR THE BOTSWANA DATA SET

No.	Class Name	Area 1	Area 2
1.	Water	270	126
2.	Hippo Grass	101	162
3.	Floodplain Grasses 1	251	158
4.	Floodplain Grasses 2	215	165
5.	Reeds	269	168
6.	Riparian	269	211
7.	Firescar	259	176
8.	Island Interior	203	154
9.	Acacia Woodlands	314	151
10.	Acacia Shrublands	248	190
11.	Acacia Grasslands	305	358
12.	Short Mopane	181	153
13.	Mixed Mopane	268	233
14.	Exposed Soils	95	89

TABLE III
CLASS NAMES AND NUMBER OF DATA POINTS FOR THE MULTITEMPORAL BOTSWANA DATA SET

No.	Class Name	May	June	July
1.	Water	118	195	185
2.	Primary Floodplain	171	192	96
3.	Riparian	177	179	164
4.	Firescar	133	196	186
5.	Island Interior	137	197	131
6.	Woodlands	149	218	169
7.	Savanna	121	189	171
8.	Short Mopane	93	166	152
9.	Exposed Soils	83	156	96

aggregated. The classes representing the various land cover types that occur in this environment are listed in Table III.

B. Experimental Methodology

For the single-image scenario, the initial labeled data (ten randomly chosen data points from each class) and the unlabeled data were extracted from the same image. For the knowledge transfer case, the labeled data are selected from a particular image subset (referred to as Area 1), and the unlabeled data are chosen from spatially or temporally distinct image data (referred to as Area 2). In this case, 75% of the Area 1 data were used for building the initial classifier, and the remaining 25% were used as the validation set. All the experiments were repeated over five different samplings of the initial labeled set.

Prior to active learning, the dimensionality of the input data was reduced using a best bases feature extractor, which reduces the feature space by recursively combining highly correlated adjacent bands. The method has been shown to be better suited for feature extraction in hyperspectral data than other methods such as Segmented Principal Components Transformation (SPCT) [36]. For the single-image scenario, the number of best bases was fixed such that there are at least five times as many initial labeled samples as the number of extracted features. In the knowledge transfer scenario, because of the availability of sufficient quantities of labeled data, from Area 1, the number of best bases was determined using a validation set. The best bases method was used as a preprocessing technique as our experiments showed this method to be less sensitive to the effect of ill-conditioned covariance matrices than the Fisher-m extractor. As detailed in Section III, the Fisher-m feature extractor was then used to obtain a discriminatory feature space from the more "stable" feature set produced by the best bases method.

The proposed active learning technique can be implemented with any classifier that makes use of estimates of *a posteriori* class probabilities for determining the decision boundaries. In our experiments, we used the ML classifier and the BHC. The ML classifier was implemented as detailed in Section III. The BHC is a multiclassifier system that was primarily developed to deal with multiclass hyperspectral data [21]. It recursively decomposes a multiclass (C-classes) problem into $(C-1)$ binary metaclass problems, which results in $(C-1)$ classifiers arranged as a binary tree. The partitioning of a parent set of classes into metaclasses is obtained through a deterministic annealing process that encourages similar classes to remain in the same partition. The metaclasses at each node of the BHC are modeled using mixtures of Gaussians, with the number of Gaussians corresponding to the number of classes at that node. Each node also has a corresponding Fisher feature extractor.

The proposed active learning method (KL-Max) detailed in Section III was implemented. Our approach was evaluated against two baseline methods, i.e., Random and Entropy. The first method chooses the data points at random, one at a time, from the unlabeled set and uses constrained EM to update the estimates of the class parameters. The entropy-based active learning approach of Roy and McCallum [26] is one of the more popular methods of active learning that make use of *a posteriori* class probabilities. As mentioned in Section II-D, this entropy-based method chooses the data points that result in an increase in the future expected entropy. Following the notation from (2) and (4), $\hat{x}$ is selected using the following equations:

$$E_{D_L^+}(\tilde{x},\tilde{y}) = \frac{1}{|D_{UL}^+|}\sum_{x\in D_{UL}^+}\sum_{y\in Y} P_{D_L^+}(y|x)\log P_{D_L^+}(y|x). \quad (5)$$

The $\hat{x}\in D_{UL}$ with the lowest expected loss is then selected for querying and is added to D_L as

$$\hat{x} = \underset{\tilde{x}\in D_{UL}}{\operatorname{argmin}} \sum_{\tilde{y}\in Y} E_{D_L^+}(\tilde{x},\tilde{y}) P_{D_L}(\tilde{y}|\tilde{x}). \quad (6)$$

To have a fair comparison, as in our proposed method, semi-supervised EM was used to estimate the parameters of the class-conditional pdfs. For reasons of computational efficiency, the new data point $\hat{x}$ was chosen from a randomly chosen subset (30 data points) of the unlabeled data for both the KL-Max and the entropy method.

V. RESULTS AND DISCUSSION

Figs. 1–4 show the learning rates of the different active learning methods. Each point on the x-axis represents the number of additional labeled samples used to train the classifier, while the y-axis represents the classification accuracies. The

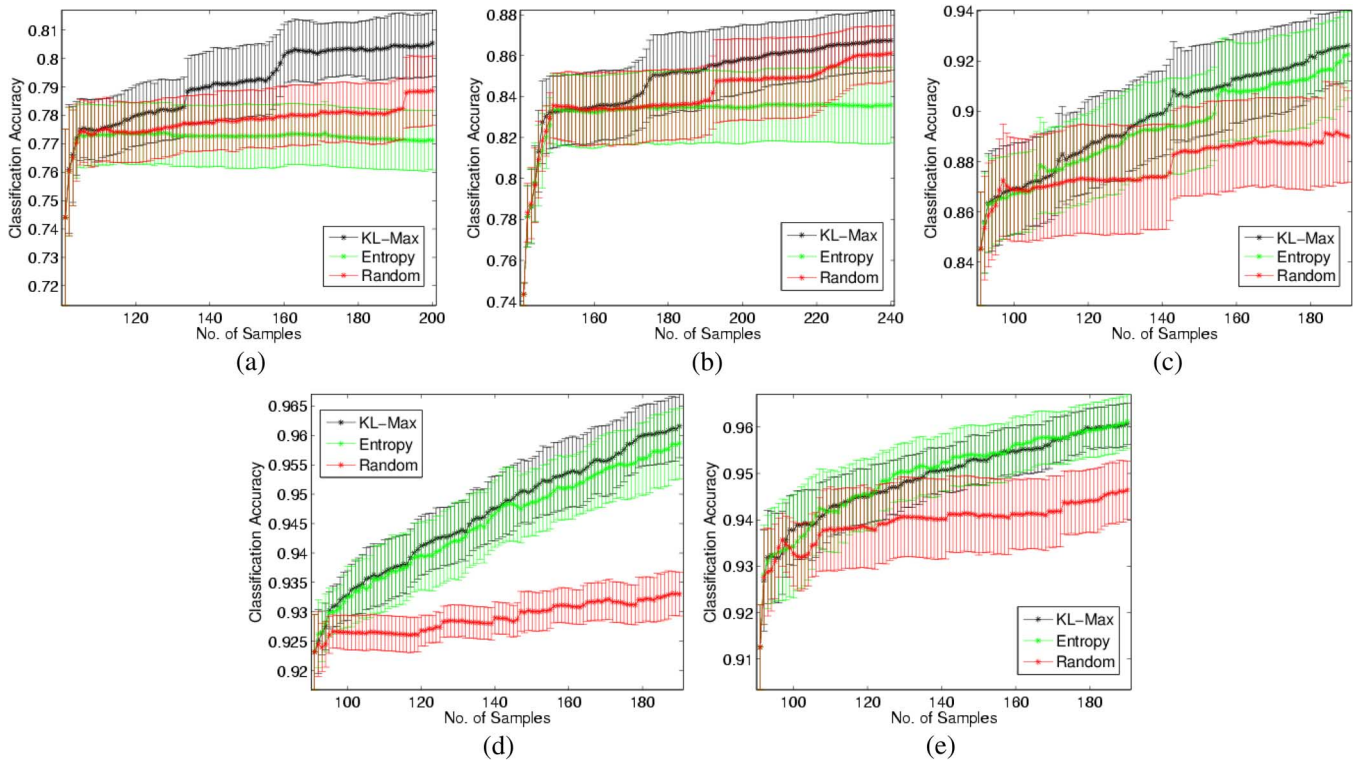

Fig. 1. Classification accuracy versus active learning iterations on a single image with the ML + EM classifier. (a) KSC. (b) Botswana. (c) May. (d) June. (e) July.

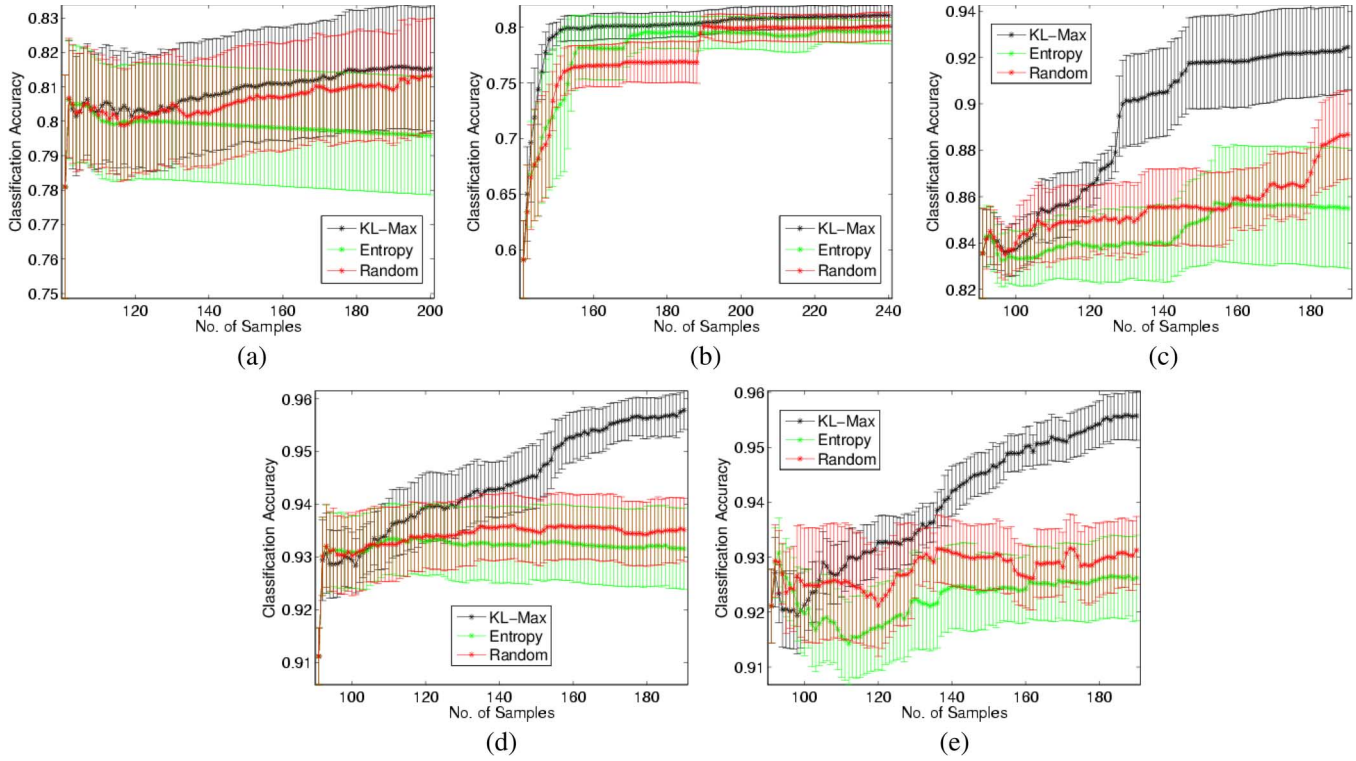

Fig. 2. Classification accuracy versus active learning iterations on a single image with the BHC + EM classifier. (a) KSC. (b) Botswana. (c) May. (d) June. (e) July.

error bars for classification accuracies were obtained using the five different samplings of the initial labeled data set, as detailed in Section IV-B.

A. *Single-Image Classification*

Figs. 1 and 2 show the learning rate curves for single-image classification over 100 active learning iterations for the different

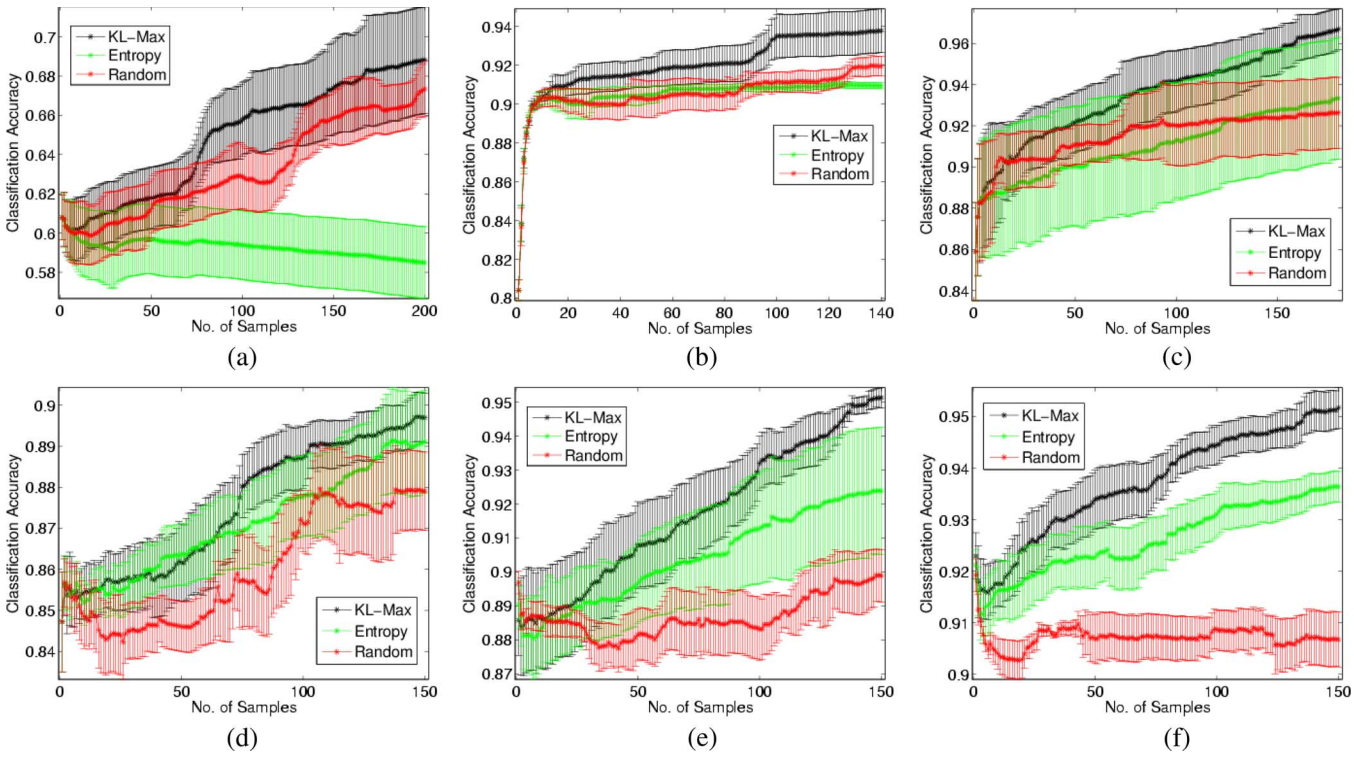

Fig. 3. Classification accuracy versus active learning iterations on spatially/temporally separate images with the ML + EM classifier. (a) Spatial KSC. (b) Spatial Botswana. (c) May to June. (d) May to July. (e) June to July. (f) May + June to July.

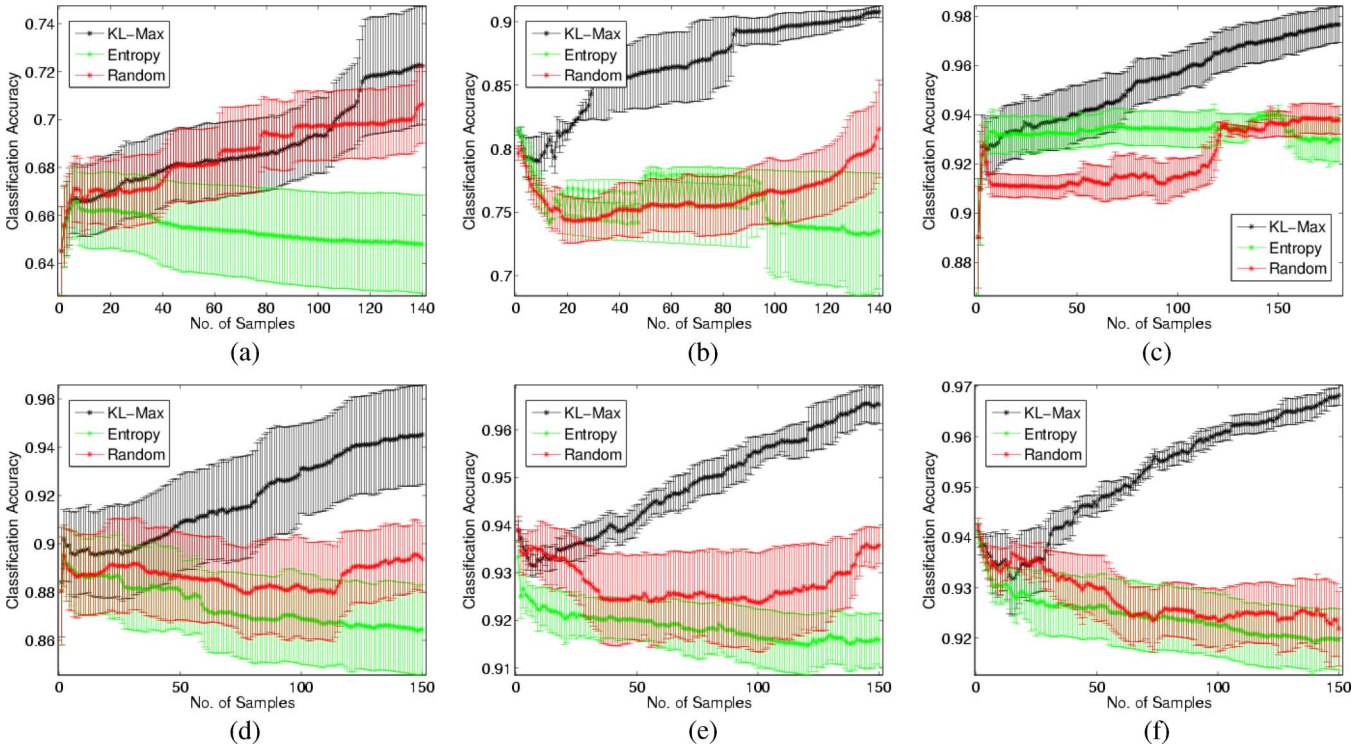

Fig. 4. Classification accuracy versus active learning iterations on spatially/temporally separate images with the BHC + EM classifier. (a) Spatial KSC. (b) Spatial Botswana. (c) May to June. (d) May to July. (e) June to July. (f) May + June to July.

data sets using the ML and BHC methods, respectively. All the active learning methods, for single-image classification, make use of an initial classifier trained using ten randomly chosen data points from each class. Thus, the x-axis for the KSC data starts at 100, the Botswana at 140, and the remaining data sets at 90. For the ML classifier, the proposed KL-Max method performs much better than the other methods on the KSC and Botswana data sets, whereas the learning rates are comparable

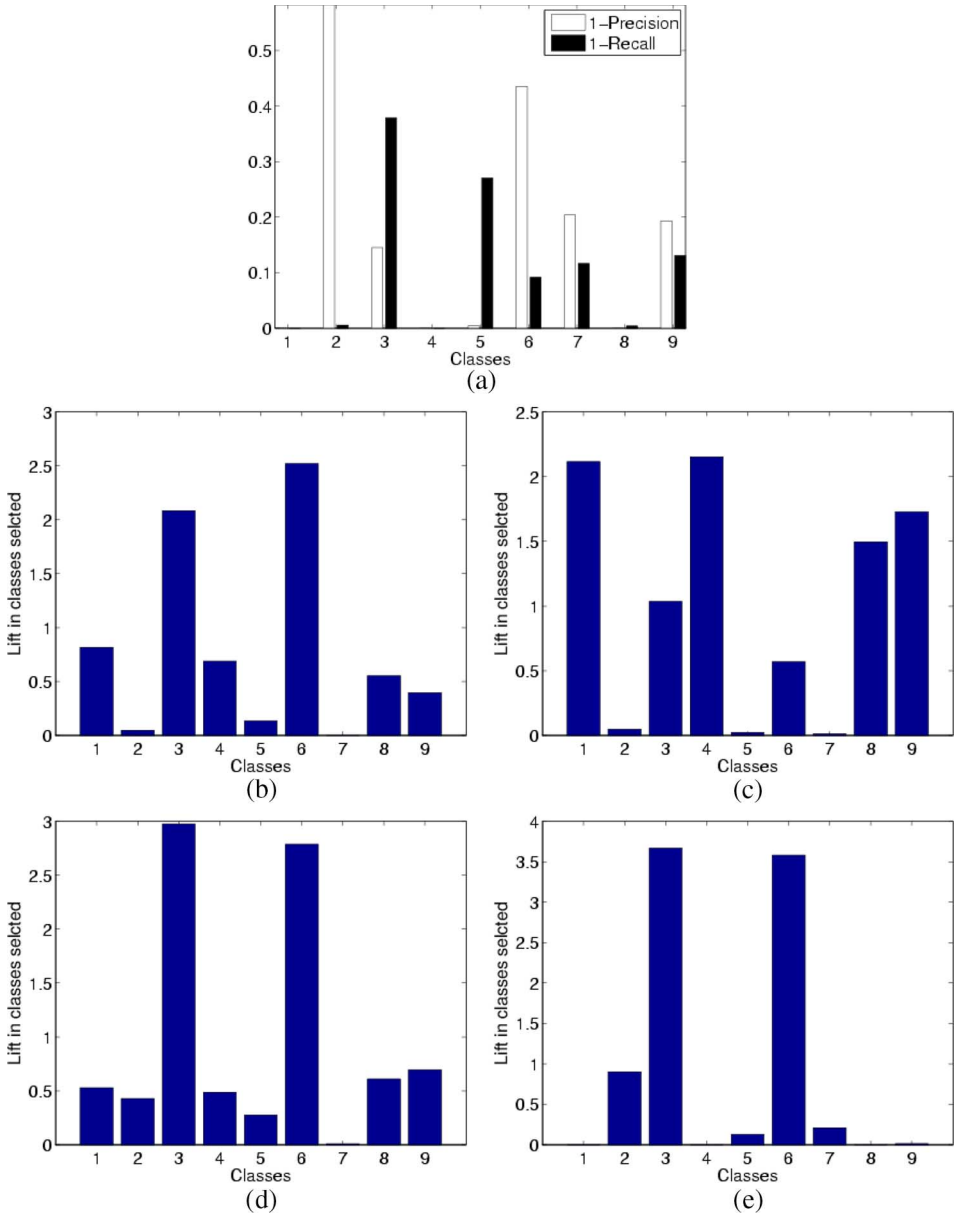

Fig. 5. Likelihood of classes being chosen by the active learning methods for the May-to-June knowledge transfer problem. (a) Per-class confusion. (b) ML entropy. (c) BHC entropy. (d) ML KL-Max. (e) BHC KL-Max.

to those of the entropy-based approach on the May, June, and July data sets. On data sets with a larger number of classes, the entropy-based method performs worse than even random selection. The poor performance of the entropy-based method could be attributed to the fact that focusing on data points that increase the future expected entropy results in skewed estimates of the class distributions.

Similar trends can be observed for the BHC method. When the proposed active learning approach was applied, both BHC and ML classifiers exhibited comparable learning behavior. However, the entropy-based method performed even worse with the BHC technique for these data sets. A possible reason for this behavior is the greater dependence of the BHC on the adequacy of the estimated class distributions. Each node in the BHC hierarchy makes use of class distribution estimates for determining the corresponding Fisher-m extractor and learning decision boundaries. Hence, skewed class distribution estimates would have an increasingly adverse effect on classification accuracies while traversing down the tree, which results in poor overall classification accuracies.

B. Knowledge Transfer

The proposed approach seems to be particularly well suited to the problem of knowledge transfer. Fig. 3 shows the learning rates for the spatially/temporally separated data sets over the active learning iterations. Note that unlike the single-image case, the x-axis for all the data sets in this case starts at zero. The KL-Max method yields higher overall classification accuracies than the other approaches. The results for the

TABLE IV
CONFUSION MATRIX FOR THE MAY-TO-JUNE KNOWLEDGE TRANSFER PROBLEM USING SEMISUPERVISED BHC

Predicted / True	1	2	3	4	5	6	7	8	9
1	195	0	0	0	0	0	0	0	0
2	0	80	0	0	74	0	0	0	38
3	0	0	53	0	0	26	0	0	0
4	0	0	0	196	0	0	0	0	0
5	0	0	0	0	196	0	0	0	1
6	0	2	87	0	0	123	5	0	1
7	0	0	0	0	1	0	150	0	38
8	0	0	0	0	0	0	0	166	0
9	0	0	0	0	2	0	27	2	126

TABLE V
CONFUSION MATRIX FOR THE MAY-TO-JUNE KNOWLEDGE TRANSFER PROBLEM USING BHC KL-MAX AFTER 180 ACTIVE LEARNING ITERATIONS

Predicted / True	1	2	3	4	5	6	7	8	9
1	195	0	0	0	0	0	0	0	0
2	0	172	0	0	2	0	0	0	0
3	0	0	94	0	0	15	0	0	0
4	0	0	0	196	0	0	0	0	0
5	0	0	0	0	194	0	0	0	0
6	0	0	15	0	0	120	0	0	0
7	0	0	0	0	1	0	185	0	0
8	0	0	0	0	0	0	0	166	0
9	0	0	0	0	2	0	4	0	152

entropy-based method are similar to those of the single-image scenario.

A comparison of the classification accuracies between the BHC using KL-Max for the single image and the knowledge transfer scenario shows that for the same amounts of available labeled data, the knowledge transfer method has higher classification accuracies than learning a classifier from scratch on the new image. For example, consider the July data set. Fig. 2(e) shows the classification accuracy when both the initial labeled data set and the unlabeled data are drawn from these data. Fig. 4(d)–(f) shows the classification accuracies for the same data set when the initial labeled data are selected from related multitemporal images, namely May and June. Fig. 2(e) shows that using 140 labeled data points from the July data results in a classification accuracy of approximately 94%. In comparison, using the knowledge in the existing June and May + June classifiers achieves the same accuracy with only about 50 data points [Fig. 4(e) and (f)]. However, classifying the July data set using the classifier trained on May data [Fig. 4(d)] requires about 120 labeled data points from the July data set to obtain the same accuracy. This is because the July data represent changes that have occurred over a two-month interval. Additionally, training data for some classes were necessarily extracted from different geographic locations in June and July due to the change in pointing angle and the advance of the flood.

A better understanding of the efficacy of the KL-Max method, for knowledge transfer, compared to the entropy-based approach, can be obtained by comparing the likelihood of the class labels chosen by these methods to the per-class confusion. In the following analysis, we measure the per-class confusion by two quantities, namely (1-precision) and (1-recall) [37]. A class with unit precision and recall values has no confusion. Hence, in our analysis, classes with high values of (1-precision) and/or (1-recall) exhibit substantial confusion, and active learning methods should be able to focus on such classes.

In the following discussion, we make use of the May-to-June knowledge transfer scenario as an illustrative example. This data set combination is representative of the remaining data sets as it exhibits the spatial and temporal variations between the May and June images. Fig. 5(a) shows the per-class confusion in classifying the June data, via semisupervision, using a BHC classifier trained on the May data. Note that this is the very first step of the active learning process. The classifier was trained using five different samplings of the May data, and the obtained results were averaged. The actual averaged confusion matrix is shown in Table IV. It can be seen that the two woodland classes, i.e., Riparian (Class 3) and Woodlands (Class 6), exhibit significant confusion. The Primary Floodplain (Class 2) class is sometimes classified as either the Island Interior (Class 5) or the Exposed Soils (Class 9) class. The Savanna (Class 7) and Exposed Soils (Class 9) land cover types also show some confusion.

Fig. 5(b)–(e) shows the lift in selecting 180 data points from each class. For each class, the lift is measured as the ratio of the number of data points chosen by the active learning method to the number of data points that would have been chosen from it by random selection. Those classes with a higher lift are more likely to have data points chosen than classes with a lower value of lift. Thus, a good active learning method should have a strong correlation between the per-class lift and the per-class confusion. It can be seen that the KL-Max method [Fig. 5(d) and (e)] has a better correlation with the per-class confusion than the entropy-based method [Fig. 5(b) and (c)]. Note that the overall best correlation is obtained with the BHC KL-Max method [Fig. 5(e)].

In addition to showing that the proposed method not only identifies the "correct" problem classes, we also show that it

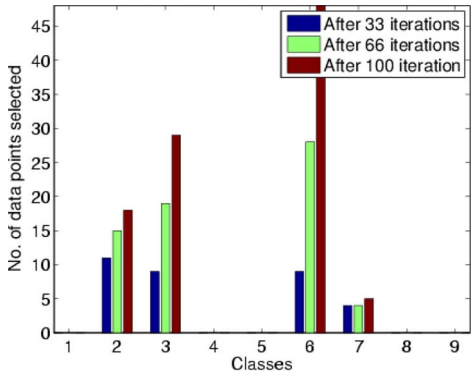

Fig. 6. Number of data points chosen from each class at different active learning iterations for the May-to-June knowledge transfer problem.

selects the most informative data points from these classes. Table V shows the averaged confusion matrix obtained by the BHC KL-Max method after 180 active learning iterations. The KL-Max method eliminates the confusion among all classes except that of Riparian (Class 3) and Woodlands (Class 6), which are both tree classes. Fig. 5(e) shows that while the KL-Max method is more likely to select data points from classes 3 and 6, the two classes continue to exhibit some confusion. To understand this behavior, consider Fig. 6 showing the distribution of class labels selected after 33, 66, and 100 iterations of a single active learning run. Note that for classes 2 and 7, the increase in the number of additional labeled data points, i.e., between 33 and 100 iterations, is far less than that of classes 3 and 6. Thus, one may conclude that for classes 2 and 7, the most informative data points are chosen early on in the active learning process, which probably is the case for classes 3 and 6 as well. However, as we force active learning to proceed, regardless of whether the estimates of class distributions change across subsequent iterations, the algorithm continues to select data points from the "more confusing" classes 3 and 6.

VI. Conclusion

We have proposed a new active learning approach for efficiently updating classifiers built from small quantities of labeled data. The principle of selecting data points that mostly change the existing belief in class distributions seems to be particularly well suited to the scenario in which the distributions of the classes show spatial (or temporal) variations. The proposed method is empirically shown to have better learning rates than choosing data points at random and an entropy-based active learning method regardless of the underlying probabilistic classifier. This paper can be expanded when more hyperspectral data are available, especially to determine the effectiveness of the active learning-based knowledge transfer framework when the spatial/temporal separation of the data sets is systematically increased.

Acknowledgment

The authors would like to thank A. Neunschwander and Y. Chen for help in preprocessing the Hyperion data.

References

[1] J. S. Pearlman, P. S. Berry, C. C. Segal, J. Shapanski, D. Beiso, and S. L. Carman, "Hyperion: A space-based imaging spectrometer," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1160–1173, Jun. 2003.

[2] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to knowledge transfer for hyperspectral data analysis," in *Proc. IGARSS*, Denver, CO, 2006, pp. 541–544.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[4] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. 16th ICML*, 1999, pp. 200–209.

[5] M. Seeger, "Learning with labeled and unlabeled data," Inst. for Adaptive Neural Comput., Univ. Edinburgh, Edinburgh, U.K., Feb. 2001. Tech. Rep.

[6] D. MacKay, "Information-based objective functions for active data selection," *Neural Comput.*, vol. 4, no. 4, pp. 590–604, Jul. 1992.

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

[8] J. T. Morgan, A. Henneguelle, M. M. Crawford, J. Ghosh, and A. Neuenschwander, "Adaptive feature spaces for land cover classification with limited ground truth," in *Multiple Classifier Systems, Lecture Notes in Computer Science*, vol. 2364, F. Roli and J. Kittler, Eds. New York: Springer-Verlag, 2002, pp. 189–200.

[9] J. A. Richards, M. M. Crawford, J. P. Kerkes, S. B. Serpico, and J. C. Tilton, "Foreword to the special issue on advances in techniques for analysis of remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 411–413, Mar. 2005.

[10] Q. Jackson and D. A. Landgrebe, "An adaptive classifier design for high-dimensional data analysis with a limited training data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 12, pp. 2264–2279, Dec. 2001.

[11] M. Dundar and D. A. Landgrebe, "A cost-effective semi-supervised classifier approach with kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 1, pp. 264–270, Jan. 2004.

[12] B. Jeon and D. A. Landgrebe, "Partially supervised classification using weighted unsupervised clustering," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 2, pp. 1073–1079, Mar. 1999.

[13] P. Mantero, G. Moser, and S. B. Serpico, "Partially supervised classification of remote sensing images through SVM-based probability density estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 559–570, Mar. 2005.

[14] P. H. Swain, "Bayesian classification in a time-varying environment," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 12, pp. 879–883, Dec. 1978.

[15] Y. Bazi, L. Bruzzone, and F. Melgani, "An approach to unsupervised change detection in multi-temporal SAR images based on the generalized Gaussian distribution," in *Proc. IGARSS*, Anchorage, AK, 2004, pp. 1402–1405.

[16] S. B. Serpico, L. Bruzzone, F. Roli, and M. A. Gomarasca, "An automatic approach for detecting land-cover transitions," in *Proc. IGARSS*, Lincoln, NE, 1996, pp. 1382–1384.

[17] B. Jeon and D. A. Landgrebe, "Decision fusion approach for multitemporal classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1227–1233, 1999.

[18] B. Jeon and D. A. Landgrebe, "Spatio temporal contextual classification of remotely sensed multispectral data," in *Proc. IEEE Int. Conf. Syst, Man, Cybern.*, Los Angeles, CA, 1990, pp. 342–344.

[19] N. Khazenie and M. M. Crawford, "Spatial–temporal autocorrelated model for contextual classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 4, pp. 529–539, Jul. 1990.

[20] L. Bruzzone and D. F. Prieto, "Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 456–460, Feb. 2001.

[21] S. Kumar, J. Ghosh, and M. M. Crawford, "Hierarchical fusion of multiple classifiers for hyperspectral data analysis," *Pattern Anal. Appl.*, vol. 5, no. 2, pp. 210–220, 2002.

[22] T. G. Dietterich and G. Bakiri, "Solving multi-class learning problems via error-correcting output codes," *J. Artif. Intell. Res.*, vol. 2, pp. 263–286, 1995.

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

[24] D. Cohn, Z. Gharamani, and M. Jordan, "Active learning with statistical models," *Artif. Intell. Res.*, vol. 4, pp. 129–145, 1996.
[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ: Wiley, 1991.
[26] N. Roy and A. K. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proc. 18th ICML*, 2001, pp. 441–448.
[27] D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proc. 17th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1994, pp. 3–12.
[28] H. S. Seung, M. Opper, and H. Smopolinsky, "Query by committee," in *Proc. 5th Annu. ACM Workshop Comput. Learning Theory*, Pittsburgh, PA, 1992, pp. 287–294.
[29] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proc. 15th ICML*, 1998, pp. 1–9.
[30] V. S. Iyengar, C. Apte, and T. Zhang, "Active learning using adaptive resampling," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery and Data Mining*, 2000, pp. 92–98.
[31] M. Saar-Tsechansky and F. J. Provost, "Active learning for class probability estimation and ranking," in *Proc. 17th Int. Joint Conf. Artif. Intell.*, 2001, pp. 911–920.
[32] P. Melville, S. M. Yang, M. Saar-Tsechansky, and R. J. Mooney, "Active learning for probability estimation using Jensen–Shannon divergence," in *Proc. 16th ECML*, 2005, pp. 268–279.
[33] A. K. McCallum and K. Nigam, "Employing EM in pool-based active learning for text classification," in *Proc. 15th ICML*, 1998, pp. 350–358.
[34] I. Muslea, S. Minton, and C. Knoblock, "Active + semi-supervised learning = robust muti-view learning," in *Proc. 19th ICML*, Sydney, Australia, 2002, pp. 435–442.
[35] Y. D. Rubinstein and T. Hastie, "Discriminative vs informative learning," in *Proc. Knowledge Discovery and Data Mining*, 1997, pp. 49–53.
[36] 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.
[37] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford Univ. Press, 1995.
[38] J. T. Morgan, "Adaptive hierarchical classifier with limited training data," Ph.D. dissertation, Dept. Mech. Eng., Univ. Texas, Austin, TX, 2002.
[39] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.
[40] [Online]. Available: www.lans.ece.utexas.edu/~rsuju/hyper.pdf

Suju Rajan received the B.E. degree in electronics and communications engineering from the University of Madras, Chennai, India, in 1997, and the M.S. and Ph.D. degrees from the University of Texas, Austin, in 2004 and 2006, respectively.

She is currently with the Department of Electrical and Computer Engineering, The University of Texas. Her research interests primarily lie in machine learning, information retrieval, and data mining.

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

Since 1988, he has been with the University of Texas (UT), Austin, where he is currently a Schlumberger Centennial Chair Professor of Electrical and Computer Engineering. He is the Founder–Director of the Intelligent Data Exploration and Analysis Lab (IDEAL). He has published more than 200 refereed papers and 30 book chapters, and co-edited 18 books. His research interests primarily lie in intelligent data analysis, data mining and web mining, adaptive multilearner systems, and their applications to a wide variety of complex engineering and AI problems.

Dr. Ghosh is the founding chair of the Data Mining Tech. Committee of the IEEE CI Society. He was the Conference Co-Chair from 1993 to 1996 and from 1999 to 2003 for the Artificial Neural Networks in Engineering (ANNIE), the Program Co-Chair for The 2006 SIAM International Conference on Data Mining, and the Conference Co-Chair of the 2007 Computational Intelligence and Data Mining. He was voted the Best Professor by the Software Engineering Executive Education Class of 2004 and has given keynote talks at several international forums. He has received 12 best paper awards, including the 2005 Best Research Paper from UT from the Co-op Society and the 1992 Darlington Award given for the best paper across all publications of the IEEE Circuits and Systems Society.

Melba M. Crawford (M'89–SM'05–F'07) received the B.S. and M.S. degrees in civil engineering from the University of Illinois, Urbana, in 1970 and 1973, respectively, and the Ph.D. degree in systems engineering from The Ohio State University, Columbus, in 1981.

From 1990 to 2005, she was a faculty member with the University of Texas, Austin. She is currently with Purdue University, West Lafayette, IN, where she is the Director of the Laboratory for Applications of Remote Sensing and the Assistant Dean for Interdisciplinary Research in the Colleges of Agriculture and Engineering. She is the holder of the Purdue Chair of Excellence in Earth Observation. She has more than 100 publications in scientific journals, conference proceedings, and technical reports, and is internationally recognized as an expert in the development of statistical methods for the analysis of hyperspectral and LIDAR remote sensing data.

Dr. Crawford was a Jefferson Senior Science Fellow at the U.S. Department of State from 2004 to 2005. She is a member of the IEEE Geoscience and Remote Sensing Society, where she is currently the Vice President for Meetings and Symposia, and an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She has 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, which received a NASA Outstanding Service Award. She is currently a member of the advisory committee to the NASA Socioeconomic Applications and Data Center at Columbia University, New York, NY.