Exploiting Class Hierarchies for Knowledge Transfer in Hyperspectral Data

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

***Abstract*—Obtaining ground truth for classification of remotely sensed data is time consuming and expensive, resulting in poorly represented signatures over large areas. In addition, the spectral signatures of a given class vary with location and/or time. Therefore, successful adaptation of a classifier designed from the available labeled data to classify new hyperspectral images acquired over other geographic locations or subsequent times is difficult, if minimal additional labeled data are available. In this paper, the binary hierarchical classifier is used to propose a knowledge transfer framework that leverages the information extracted from the existing labeled data to classify spatially separate and multitemporal test data. Experimental results show that in the absence of any labeled data in the new area, the approach is better than a direct application of the original classifier on the new data. Moreover, when small amounts of the labeled data are available from the new area, the framework offers further improvements through semisupervised learning mechanisms and compares favorably with previously proposed methods.**

***Index Terms*—Hierarchical classifier, knowledge transfer, multitemporal data, semisupervised classifiers, spatially separate data.**

I. Introduction

A COMMON application of hyperspectral imaging involves mapping spectral signatures in the images to specific land-cover types. While hyperspectral data are now readily available, obtaining reliable and accurate class labels for each "pixel" is a nontrivial task involving expensive field campaigns and time-consuming manual interpretation of imagery. Typically, the labeled ground-truth data are acquired over spatially contiguous sites that are easily accessible. Such "spatially localized" data are then used to classify the entire hyperspectral image including those regions, from which no labeled data were obtained [1], [2]. Implicit in this method of classification is the assumption that the spectral signatures of each land-cover type do not exhibit substantial spatial (or temporal) variations. However, factors such as soil composition, topographic variations, and local atmospheric condition alter the spectral characteristics measured at the sensor, even though they correspond to the same land-cover type, from one region to another. Moreover, airborne hyperspectral data for an area of interest are typically obtained over multiple flightlines. In such cases, factors such as bidirectional reflectance can cause further variations in the class-specific spectral signatures. Hence, the naive use of a classifier which is trained on the available ground-truth data from one region on data that are from spatially or temporally different areas without accounting for the variability of the class signatures, will result in poor classification accuracies [3], [4]. Theoretically, an ideal approach would be to pool the data from all regions of interest to train a classifier that performs well over all the regions. However, researchers are typically unable to follow this path.

In this paper, we study a more feasible middle ground of exploiting certain properties of a classifier which is trained using the data acquired from one area to help classify the data obtained from spatially and temporally different areas. Thus, a key idea in our framework is to exploit the contextual information in existing classifiers for rapidly constructing a new classifier for a new but related problem, even with little additional labeled data. Specifically, we use a multiclassifier system called the binary hierarchical classifier (BHC) [5] 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, facilitates the transfer of 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 in order to better reflect the statistics of the data from new areas. Besides the unsupervised setting, the framework presented here can also be utilized when very small quantities of the labeled data are available from the spatially or temporally separate areas. We present results of experiments that demonstrate the advantages of our proposed framework over other powerful multiclassifier systems, such as the error correcting output code (ECOC) [6], for the purposes of knowledge transfer in hyperspectral data.

II. Related Work

This paper focuses on the problem of adapting a hyperspectral classifier to generate land-cover labels for future incoming data from a spatially or temporally different image. The noteworthy characteristics of this classification task are:

1) availability of large quantities of unlabeled data;
2) possibility of a population drift in the unlabeled data.

A. Semisupervised Classification and Transfer Learning

Several machine-learning approaches have been proposed to deal with certain aspects of the problem stated above.

Incorporating unlabeled data into the classification task: Given a mixture of labeled and unlabeled data, the semisupervised classification algorithms [7] try to improve the classification accuracy by making use of unlabeled data to obtain better classification boundaries. Semisupervised methods that

Manuscript received December 6, 2005; revised March 4, 2006. 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 School of Civil Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: mcrawford@purdue.edu).

Digital Object Identifier 10.1109/TGRS.2006.878442

make use of expectation maximization (EM) have met with a considerable success in a number of domains, especially that of text data analysis and remote sensing. These algorithms attempt to maximize the joint log-likelihood of both the labeled and the unlabeled data by iteratively reestimating the conditional distribution of a hidden variable, indicating the true class associated with the unlabeled data. It is important to note that the standard formulation for semisupervised classification techniques assumes that both the labeled and unlabeled data have a common underlying distribution. This assumption is violated for the application scenario considered in this paper, since it is likely that the statistics of the unlabeled data differ somewhat from that of the previously acquired labeled data, i.e., there is a population drift [8].

Classification in the presence of population drift: Several attempts at dealing with the problem of changing populations have borrowed ideas from the field of online learning or incremental learning. Most online algorithms are designed as feedback systems. It is assumed that there is a steady stream of objects to be classified, the true labels of which are revealed immediately after classification. Prior knowledge about the type of the population drift (random/gradual/sudden) is then used to update the existing classifier at regular intervals [8], [9]. The online algorithms attempt to minimize the cumulative number of errors made, such that the resulting classifier does not perform much worse than a classifier trained on the same data in batch mode. A detailed review of several online-learning algorithms is provided in [10].

Adapting online methods to deal with the problem of population drift typically involves maintaining a window of the incoming training samples. At appropriate intervals, the set of hypotheses is either retrained, or some of the outdated hypotheses are removed and replaced with those that are more consistent with the recently seen observations [9]. While remotely sensed data obtained over extensive regions (or different times) also exhibit the problem of "population drift," unlike the online frameworks, one does not have access to a streaming set of labeled data samples.

Knowledge transfer and reuse: The vast majority of works in machine learning and data mining focuses on solving a specific classification task, which is isolated from other tasks. However, in practice, one is often faced by a series of (possibly related) tasks, or a task whose nature changes substantially with time. Existing techniques require large quantities of labeled data to be able to deal effectively with the changes in a classification task. The question then is, can one transfer the knowledge in a previously learned classifier to better tackle the latest classification task, instead of totally reinitiating the analysis, as is usually done?

In the mid-1990s, a first generation of approaches that involved explicit knowledge transfer/reuse emerged, under labels such as "knowledge transfer," "learning to learn," context sensitivity/drift, and "lifelong learning" [11]–[13]. For instance, several researchers attempted to directly reuse the internal state information from classifiers under the belief that related classification tasks may benefit from common internal features. Some of the other approaches of knowledge reuse include the use of supplemental training examples or historical training information, such as learning rates, reusing the labels produced by original classifiers to improve the generalization performance on a new classifier for a different but related task, and multitask learning neural networks that are trained simultaneously to perform several related classification tasks.

B. Semisupervised Learning and Knowledge Transfer for Remote Sensing Applications

The advantages of using unlabeled data to aid the classification process in the domain of remote sensing data was first identified and exploited by Shahshahani and Landgrebe [1]. In this work, they made use of the unlabeled data via EM to obtain better estimates of class-specific parameters. It was shown that using unlabeled data enhanced the performance of the maximum *a posteriori* probability (MAP) 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 [14], [15]. In these methods, the available labeled data are first used to train a supervised classifier to obtain tentative labels for the unlabeled data. Semilabeled data, thus obtained, are then used to retrain the existing classifier, and the process is iterated until convergence. Note that these methods assume that the labeled and the unlabeled data are drawn from the same distribution. In other words, the estimated class parameters are considered unreliable because of the nonavailability of the labeled data and not because of changes in the underlying data distribution.

Besides the typical semisupervised setting, unlabeled data have also been utilized for "partially supervised classification" [16], [17]. In partially supervised classification problems, the 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.* [17] attempt to model the distribution of the class of interest and automatically determine a suitable "acceptance probability," Jeon and Landgrebe [16] make use of the unlabeled data while learning a maximum-likelihood (ML) classifier to determine whether a data point is of interest or not.

While all these methods deal with the data obtained from the same image, the possibility that the class label of a pixel could change with time was first explored in [18]. In this work, the joint probabilities of all possible combinations of classes between the multitemporal images were estimated and used in the classification rule. The proposed "multitemporal cascade classifier," however requires the labeled data from all the images of interest. More recently, unsupervised algorithms, which can automatically detect whether a particular pixel in multitemporal images has changed have also been proposed [19]. Besides algorithms for change detection, supervised algorithms which automatically try to model the class transitions in multitemporal images have also been developed [20]. Another supervised attempt at classifying multitemporal images involves building a local classifier for each image, the decisions of which are then combined, either via a joint likelihood-based rule or a weighted majority decision rule that takes into account the reliabilities of the data sets and that of the individual classes, to yield a "global" decision rule for the unlabeled data [21]. Similarly, other spatial-temporal methods utilize the temporal correlation of the classes between images to help improve the classification accuracy [22], [23].

A pioneering attempt at unsupervised knowledge transfer for multitemporal remote sensing images was made in [3]. In this work, 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 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. Experimental results later revealed that the simple ML-based knowledge transfer did not perform as expected for "complex" data sets [24]. The authors therefore recommend using an ensemble of "complementary" classifiers. In particular, the ML classifier and two radial basis function (RBF) neural networks were first trained on the labeled data from t_1. The classifiers, thus obtained, were then updated using the unlabeled data from t_2 via EM. The results of the ensemble were then combined either by a majority voting, a Bayesian combination method, or by the MAP rule. For these experiments, the ensemble yielded higher classification accuracies than the EM-updated ML classifier.

While Bruzzone *et al.* [3], [24] demonstrate the advantage of using previously acquired knowledge in classifying a novel image, the amount of knowledge transferred was restricted by the classifiers under consideration, namely the ML and the RBF neural net classifier. The only knowledge from the training data that was transferred in this framework was the set of estimates of the parameters of the class distributions modeled as Gaussians. Using other classifier systems might enable one to extract and transfer more information from the available training data. It is in this context that we propose using the BHC as the classifier in our knowledge transfer framework.

C. BHC

The BHC [5] is a multiclassifier system that was developed primarily to deal with multiclass hyperspectral data. The BHC involves recursively decomposing a multiclass (C-classes) problem into $(C-1)$ binary metaclass problems, resulting in $(C-1)$ classifiers arranged as a binary tree. The given set of classes is first partitioned into two disjoint metaclasses, and each metaclass, thus obtained, is partitioned recursively until it contains only one of the original classes. The number of leaf nodes in the tree is thus equal to the number of classes in the output space. The partitioning of a parent set of classes into metaclasses is not arbitrary, but is obtained through a deterministic annealing process, which encourages similar classes to remain in the same partition [5]. To combat the small sample size problem in analyzing the hyperspectral data, the dimensionality of the feature space is reduced by recursively combining highly correlated adjacent bands [25]. This "best bases" method of a feature extraction, which makes use of the class information as the correlation between bands, varies among the classes, thereby yielding an interpretable feature space.

The BHC offers comparable classification accuracies to those achieved by other multiclassifier systems such as the ECOC [26], but it seems more suitable for knowledge transfer than other alternatives as it reveals additional knowledge. The hierarchy of classes, for instance, might be useful as the relationships between classes in one area might still hold in another area. Further, since the best bases feature-extraction method makes use of class-specific information in determining the set of adjacent bands that are to be merged, this information can also be exploited in the new area. Finally, the Fisher discriminant makes use of both within-class and between-class covariances, which can also be helpful, as we might expect similar correlations between the classes in the new area.

The generalization ability of the BHC used within a random forest framework for the analysis of spatially separate data was studied in [4]. The random forest technique improves performance but does not explicitly transfer any knowledge from an existing forest of BHCs to the new classification problem. The first attempt at transferring the information from an existing BHC to classify a new region with no ground-truth data is described in [27]. The proposed method made use of the Fisher discriminant associated with each metaclass pair to project new unlabeled data into the corresponding Fisher space. The projected data were then clustered using the k-means algorithm. Finally, the resulting clusters were assigned to the metaclasses, such that the distances between the cluster centers and the corresponding metaclass means were minimized. The resulting "pseudo-labeled" data were used to update the parameter estimates in the BHC tree. It was found that while the updated classifier improved the classification accuracies on one hyperspectral data set, its performance on another data set was slightly worse than the naive application of the existing BHC to the unlabeled data.

III. Knowledge Transfer Framework

Let us assume that we have the hyperspectral data from two spatially (or temporally) different areas, Area 1 and 2. Let us also suppose that for Area 1, there is an adequate amount of labeled data to build a supervised classifier. We first consider the situation where all the data from Area 2 are unlabeled (unsupervised case). Subsequently, the impact on design and performance of the proposed framework is studied when labels are provided for a small part of the data from Area 2 (semisupervised case).

A. Unsupervised Case

In the absence of any labeled data from Area 2, the first step in the knowledge transfer framework is to use the training data from Area 1 to generate the corresponding BHC tree. We then attempt to transfer the knowledge in this BHC to Area 2.

Our first approach was to use the hierarchy of the classes and the best bases feature extractors of the BHC classifier built on the Area 1 data, but modify the Fisher feature extractors and the binary classifiers to account for the changed statistics of the spatially separate data. This was achieved via the EM framework, in which the training data were used to initialize the EM algorithm, and the data from Area 2 were treated as unlabeled. At each node, the corresponding Fisher extractor was first used to project the data from both Area 1 and Area 2 into a reduced dimensionality space. The metaclasses at

the node were modeled using mixtures of Gaussians, with the number of Gaussians corresponding to the number of classes at that node. The initial parameters of the Gaussians were estimated using the corresponding class data from Area 1. In the E-step of the algorithm, the Gaussians were used to determine the posterior probabilities of the Area 2 data. The probabilities, thus estimated, were then used to update the parameters of the Gaussians (M-step). EM iterations were performed until the average change in the posterior probabilities between two iterations was smaller than a specified threshold [3]. A new Fisher feature extractor was also computed for each EM iteration, which is based on the statistics of the metaclasses at that iteration. The updated extractor was then used to project the data into the corresponding Fisher space prior to the estimation of the class-conditional pdfs.

Analysis of the results showed that while this approach yielded somewhat higher overall classification accuracies than a direct application of the original classifier, the errors were mostly concentrated in a few classes. A closer inspection revealed that the spectral signatures of these classes had changed sufficiently for them to be grouped differently in the BHC hierarchies, if there had been adequate amounts of labeled data from Area 2. This suggested that we should have obtained multiple trees from Area 1, such that some of them would be more suitable for the new area.

Thus, our second approach was to introduce randomization into the structure of the BHC tree. The design space for the BHC offers many possibilities for randomizing the tree structure. In our earlier work [28], we generated randomized BHC trees by varying factors such as the percentage of the available training data, the number of features selected at each node, class priors, and by randomly switching the class labels for a small percentage of the labeled data points. In this paper, randomized BHC trees were generated by choosing an internal node of the tree and randomly interchanging the classes drawn from its right and left children. The corresponding feature extractors and classifiers at that node (and its children) were then updated to reflect the perturbation. Note that in the absence of any labeled data from Area 2, there is no way to evaluate which of the randomly generated BHC trees best suits the spatially/temporally different data. Hence, we can only generate an ensemble of classifiers using the training data, hoping that the ensemble contains some classifiers that are better suited to Area 2.

The key to the success of an ensemble of classifiers is choosing the classifiers that make independent errors. If the classifiers are not independent, the ensemble might actually perform worse than the best member of the ensemble. Hence, a number of measures of diversity have been proposed to choose a good subset of classifiers [29]. Of the ten diversity measures studied, the authors recommend the Q_{av}, the ρ_{av}, and the κ measures for their easy interpretability. They further promote the Q-diversity measure because of its relationship with the majority vote of an ensemble and its ease of calculation. Hence, we made use of the Q-diversity measure in our earlier study [28]. However, on experimenting with the κ measure [30], we found that it yielded a comparable, if not better, performance in the sense of resulting overall classification accuracy than that of the Q measure. Further, unlike the Q-diversity measure, the κ measure does not require access to any labeled data. Hence, in this paper, the κ-diversity measure, which indicates the degree of disagreement between a pair of classifiers, was used to ensure the diversity of our classifier ensemble.

The data from Area 2 were labeled using each tree in the classifier ensemble, and these labels were then used to obtain the κ measure between each pair of classifiers. The classification results of a smaller set of classifiers with the lowest average pairwise κ measure (i.e., higher diversity) were then combined via a simple majority voting.

B. Semisupervised Case

If small amounts of labeled data are available, knowledge transfer mechanisms can improve classification accuracies, especially if they exploit the added information. In this section, we generalize both knowledge transfer methods in order to leverage the labeled data and determine how much labeled data are required from the spatially separate area before the advantages of transferring information from the original solution are no longer realized.

The ensemble-based approach was modified in two stages. First, after the set of classifiers was pruned to improve the diversity of the ensemble by using the κ-diversity measure, we further pruned the remaining set of classifiers to include only those which had yielded higher classification accuracies on the labeled data. A scheme similar to the online weighted majority algorithm [31], which assigns all classifiers a weight, was then used to weight the different classifiers. Prior to learning, the weights of all the classifiers are equal. As each data sample is presented to the ensemble, a classifier's weight is subsequently reduced multiplicatively, if that example is misclassified. For each new example, the ensemble then returns the class with the maximum total weighted vote over all the classifiers. Thus, the algorithm used for computing the class label predicted by the BHC ensemble is as follows.

Weighted majority vote for BHC ensemble

1) Initialize the weights $w_1, \ldots, w_n$ of all n BHCs to 1.
2) For each labeled data point, let $y_1, \ldots, y_n$ be the set of class labels predicted by the BHCs.
3) Output class h_i if $\forall h_j \neq h_i$, $j = 1, \ldots, m$, where m is the number of classes

$$\sum_{k=1; y_k == h_i}^{n} w_k \geq \sum_{k=1; y_k == h_j}^{n} w_k.$$

4) On observing the correct class label, if h_i is wrong, then multiply the weight of each incorrect BHC by 0.5; else if h_i is correct, do not modify the weights.

At the end of this learning, the "winnowing property" of the weighted majority scheme assigns lower weights to those classifiers with poorer classification accuracies on the incoming data. Thus, by reducing the contribution of the inaccurate classifiers to the final decision, the voting scheme ensures that the performance of the ensemble is not much worse than that of the best individual predictor, regardless of the dependence between the members of the ensemble [31]. For the semisupervised implementation, the EM-based method was modified to perform a constrained EM. Here, the E-step only updates the posterior probabilities (memberships) for the unlabeled data while fixing the memberships of the labeled instances according

to the known class assignments [32]. The labeled data were also used to initialize the mean vectors and the covariance matrices of the metaclasses at the nodes of the binary trees in the κ-diversity measure pruned ensemble. The labeled and the unlabeled data from Area 2 were then used for constrained EM while updating the Fisher extractors in each of the binary trees. The classification results of the resulting ensemble were then combined using the weighted majority algorithm as detailed previously.

IV. Experimental Evaluation

In this section, we provide empirical evidence that in the absence of labeled data from the spatially/temporally separate area, using knowledge transfer is better than the direct application of existing classifiers to this new area. We also present results showing that with small amounts of the labeled data from the new areas, our framework yields higher overall accuracies for our experiments than the current state-of-the-art ECOC multiclassifier system [6] with support vector machines (SVMs) [33] as the binary classifiers. Besides the ECOC classifier, we also compare our framework with two EM-based ML (ML-EM) techniques. The first ML-EM classifier is the unsupervised approach suggested in [1]. The second method is the knowledge transfer method proposed in [3], which we refer to as a seeded ML-EM, since it uses the Area 1 data only to initialize the Gaussians prior to performing the EM iterations. The parameters of the Gaussians and the Fisher feature extractors are then updated using the unlabeled data (and, if available, labeled data) from Area 2 via EM.

A. Data Sets

The knowledge transfer approaches described above were tested on the hyperspectral data sets obtained from two sites: NASA's John F. Kennedy Space Center (KSC), Florida [27] and the Okavango Delta, Botswana [4].

1) KSC: The NASA Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) acquired the data over the KSC on March 23, 1996. AVIRIS acquires data in 242 bands of 10-nm width from 400–2500 nm. The KSC data, which are collected from an altitude of approximately 20 km, have a spatial resolution of 18 m. Removal of noisy and water absorption bands resulted in 176 candidate features. 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. 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 512×614 spatially removed test set (Area 2) is a different subset of the flight line than the 512×614 data set from Area 1 [34]. While the number of classes in the two regions differs, we restrict ourselves to those classes that are present in both regions. Details of the ten land-cover classes considered in the KSC area are in Table I.

2) Botswana: This 1476×256 pixel study area is located in the Okavango Delta, Botswana, and has 14 different land-cover types consisting of seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the delta. Data from this region were obtained by the NASA Earth Observing 1 (EO-1) satellite for the calibration/validation portion of the mission in 2001. The Hyperion sensor on EO-1 acquires data at 30-m pixel resolution over a 7.7-km strip in 242 bands, covering the 400–2500-nm portion of the spectrum in 10-nm windows. Uncalibrated and noisy bands that cover water absorption features were removed, resulting in 145 features. The land-cover classes in this study were chosen to reflect the impact of flooding on vegetation in the study area. Training data were selected manually using a combination of global positioning system (GPS)-located vegetation surveys, aerial photography from the Aquarap (2000) project, and 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 [34]. Details of the Botswana data are listed in Table II.

Multitemporal data: In order 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. While the May scene is characterized by the onset of the annual flooding cycle and some newly burned areas, the progression of the flood and the corresponding vegetation responses are seen in the June and July data. The Botswana data acquired in May had 14 classes, but only nine classes were identified for 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 were aggregated in some finer grained classes. The classes representing the various land-cover types that occur in this environment are listed in Table III.

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

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	133
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

B. Experimental Methodology

In all the data sets, the labeled data (Area 1) were subsampled, such that 75% of the data were used for training and 25% as the test set. For both cases, a second test set was also acquired from the spatially/temporally separate region (Area 2). Since the Area 2 test set was from a different geographic location, or was obtained at a different time, factors such as localized geomorphology, meteorology, and atmospheric conditions as well as changes in bidirectional reflectance and plant physiology resulted in different hyperspectral signatures. Along with the changes in the *a priori* probabilities of the land-cover classes, these data provide an ideal setting to test the knowledge transfer framework.

For our experiments, we used a BHC based on the Fisher-m feature extractor, and the posterior probabilities were obtained by soft combining. Adjacent hyperspectral bands that were highly correlated were merged using the best bases feature-extraction technique [25] prior to applying the Fisher feature extractor. Merging was performed until the ratio of the training samples to the number of dimensions was at least five at each node of the classifier tree [35]. For both the unsupervised and the semisupervised cases, the classification accuracies were obtained by averaging over five different samplings of the training data (from Area 1) or the labeled Area 2 data, respectively.

The ensemble of the BHC trees was generated by switching randomly chosen sibling classes of the original BHC tree. The feature extractors and the classifiers of the corresponding node and that of its children were then updated to reflect the perturbation. One hundred different randomized BHC trees were generated. The κ-diversity measure was then used to prune the ensemble, such that the final ensemble contained the ten classifiers with the lowest average pairwise κ measure. Earlier experiments with a larger pool of randomized trees, from which the ten most diverse classifiers were chosen, yielded similar results [28]. Note that there are no well-defined methods for determining the number of classifiers to be used in an ensemble. For our purposes, we found that generating 100 randomized trees formed an adequate initial pool of classifiers, from which, we selected ten.

For the semisupervised scenario, using very small amounts of labeled data to estimate the class covariance matrices resulted in ill-conditioned matrices. In the knowledge transfer framework, the class covariance matrices were initially stabilized by pooling the corresponding training data from Area 1, and the labeled data from Area 2 to estimate the covariance matrices. Similarly, while building a new BHC using the available Area 2 data, the estimates of the ill-conditioned class covariance matrices at a particular node were stabilized by using the data points associated with the corresponding parent node.

Both the ML-EM classifiers were modeled using a multivariate Gaussian for each class. As in the case of the BHC-based knowledge transfer, the best bases feature extractor and the Fisher discriminant were used to reduce the dimensionality of the input data. The number of best bases was determined by using a validation set from the Area 1 training data. For the unsupervised case, the best bases feature extractor was transferred from Area 1 to Area 2. Area 2 data were treated as the unlabeled data, and EM iterations were performed as detailed in Section III-A. For the semisupervised scenario, the constrained EM was used to update the parameters of the Gaussians as well as the Fisher discriminant.

For the ECOC-SVM systems, the guidelines provided in [6] were used to generate the appropriate code matrices. In our paper, we used the following:

1) dense random code method of [36] for the KSC and the multitemporal data sets;
2) BCH code matrix from [37] for the spatially separate Botswana data set.

SVMs with Gaussian kernels were trained for each binary problem induced by the code matrix [6]. The SVM classifiers were implemented in MATLAB using the package provided in [38]. Prior to SVM classification, each feature in the training data was normalized to have a zero mean and unit variance. The features of the corresponding test set were also scaled with the means and variances computed from the training data. The parameters (Gaussian kernel width and the upper bound on the coefficients of the support vectors, "C") of each SVM were identified by threefold cross validation, using 40% of the available training data as the validation set. Different values for the Gaussian kernel widths were evaluated empirically, and the parameter, which had the least classification error over the three validation sets, was finally used. Having fixed the kernel width, a similar process was used to tune the "C" parameter.

C. Results and Discussion

Unsupervised case: First, the BHC, the ECOC-SVM, and the BHC ensemble built on the training data from Area 1 were used without any modification to classify the data from Area 2. Tables IV and V contain the overall classification accuracies, along with the standard deviations of the overall accuracies, which are obtained by the baseline and the knowledge transfer approaches on the Area 2 data.

As a frame of reference, for the spatially separate data, the classification accuracies on the Area 1 test set for the BHC, ECOC-SVM, and ML + EM are 93.05%($\pm$1.17), 93%($\pm$1.03), and 89.15%($\pm$1.28) for the KSC data set. For the Botswana data set, the corresponding classification accuracies are 94.52%($\pm$0.79), 95.63%($\pm$0.95), and 93.57%($\pm$1.63), respectively. The substantial reduction in overall classification accuracies when the original classifiers were applied to spatially separate test cases shows that there is a significant difference between Area 1 and Area 2. Because of greater homogeneity within the scene, the Botswana data set benefits much more from the information in Area 1 than the KSC (Table IV). The greater disparity in the spectral signatures of the classes

TABLE IV
AVERAGE UNSUPERVISED CLASSIFICATION ACCURACIES FOR THE SPATIALLY SEPARATE TEST SETS

	Baselines				Knowledge Transfer Approaches		
Name	Orig. BHC	Orig. ECOC +SVM	Ensemble BHC +Maj. Vote	ML+EM	Orig. BHC +EM	Ensemble BHC +EM +Maj. Vote	Seeded ML+EM
KSC	61.47 (0.32)	64.27 (0.21)	65.34 (0.36)	63.39 (0.50)	62.50 (0.72)	67.92 (0.80)	64.03 (0.75)
Botswana	74.12 (1.2)	75.22(0.29)	73.69 (0.90)	80.11 (0.37)	82.30 (0.76)	82.95 (1.2)	84.42 (0.97)

TABLE V
AVERAGE UNSUPERVISED CLASSIFICATION ACCURACIES FOR THE MULTITEMPORAL TEST SETS

	Baselines				Knowledge Transfer Approaches		
Name	Orig. BHC	Orig. ECOC +SVM	Ensemble BHC +Maj. Vote	ML+EM	Orig. BHC + EM	Ensemble BHC +EM +Maj. Vote	Seeded Ml+EM
Botswana May to June	49.41 (1.4)	69.09 (0.31)	49.14 (1.31)	60.95 (8.9)	71.27 (3.6)	73.11 (2.1)	57.39 (11.1)
Botswana May to July	71.28 (1.89)	73.10 (0.43)	72.04 (1.58)	63.81 (8.72)	79.33 (2.54)	78.98 (2.24)	62.74 (7.77)
Botswana June to July	82.71 (0.21)	85.90 (0.21)	83.96 (0.37)	90.53 (1.10)	86.71 (0.20)	86.54 (1.11)	89.97 (0.99)
Botswana May+June to July	83.71 (1.06)	89.72 (0.09)	86.63 (0.86)	91.9 (0.21)	90.70 (0.29)	91.05 (0.35)	91.07 (0.60)

between the two areas in the KSC data set limits the amount of knowledge that can be transferred from one area to another. The changes in the class spectral signatures (thereby, class hierarchies) between the two areas also explain the greater gains offered by the ensemble compared to EM-based methods for this data set.

For the multitemporal images, it can be seen that the Botswana data sets benefit from the knowledge in Area 1 data (Table V). The superior classification performance of the May+June classifier on the July data set shows the utility of the knowledge transfer framework in a multitemporal scenario. Note that the May training data were not spatially colocated with the June and July data, as the scene coverage was somewhat different. The sensitivity of the EM algorithm to its initialization is clearly seen in the large standard deviations associated with the classification accuracies of the ML+EM classifiers for this dataset. However, using the hierarchy along with the EM helps reduce the effect of poor initialization.

Semisupervised case: Fig. 1 shows the learning curves for the KSC and the Botswana data sets when the labeled data are available from Area 2. The error bars denote the standard deviation of the accuracies measured over five random samplings of the labeled data. A scaled version of the learning curves for the top five techniques for the May to June and May to July temporal datasets is shown in Figs. 2 and 3. It can be observed from Fig. 1 that the ensemble with the weighted majority vote does not offer any advantage over the other classification systems, especially when there is an adequate amount of labeled data. For the KSC and the Botswana data sets, examination of the weights assigned to the classifiers of the ensemble showed that when the number of labeled samples per class (> 10) was high, the classifiers in the ensemble had almost equal weights. Hence, the accuracy of the ensemble was limited by the classification accuracies of its constituent classifiers.

Using the data (labeled and unlabeled) from Area 2 with the EM to update the statistics of the classifiers improved the classification accuracy of the original BHC in all cases. Note that for small amounts of labeled data using the knowledge from the old area actually yield greater gains in accuracy than the new BHC (built from the available labeled data).

By adapting the BHC ensemble components via constrained EM, some members became more effective for the new area. The weighted majority algorithm was then able to exploit this differentiation to produce a knowledge transfer framework that proved a clear winner for small amounts of labeled data (Fig. 1). While the ML-EM classifiers appear to be quite competitive, the real benefit of the knowledge transfer framework can be seen in the harder KSC and the multitemporal Botswana datasets. These results show that the class hierarchy learned on a related task is a useful tool for knowledge transfer, especially when the distribution of data changes significantly between the tasks. As more labeled data become available from Area 2, new classifiers trained on that data will eventually match or surpass the performance of the updated classifiers from Area 1. The amount of the labeled data from Area 2, which is required for this crossover, is surprisingly large, thereby validating the efficacy of our proposed technique.

V. CONCLUSION

We initially believed that the original BHC framework would be adequate for knowledge transfer, since it provides not only a class hierarchy but also the feature extractors that are suitable for resolving the dichotomies involved at the different stages of the hierarchy. In particular, it should be more effective than alternative classifiers, including the ML-based approach investigated earlier. However, in this application, the data characteristics change fairly substantially from area to area, demanding more extensive adjustments. The best suited class hierarchies as well as the most appropriate feature extractors change at least incrementally as one moves to a new area. We were able to cater to both these needs by 1) using the weighted majority combining approach on an ensemble of trees, so that trees, which are more suitable for the new area, get higher weights, and 2) using constrained semisupervised EM that can adjust the feature spaces as well as classification boundaries based on both

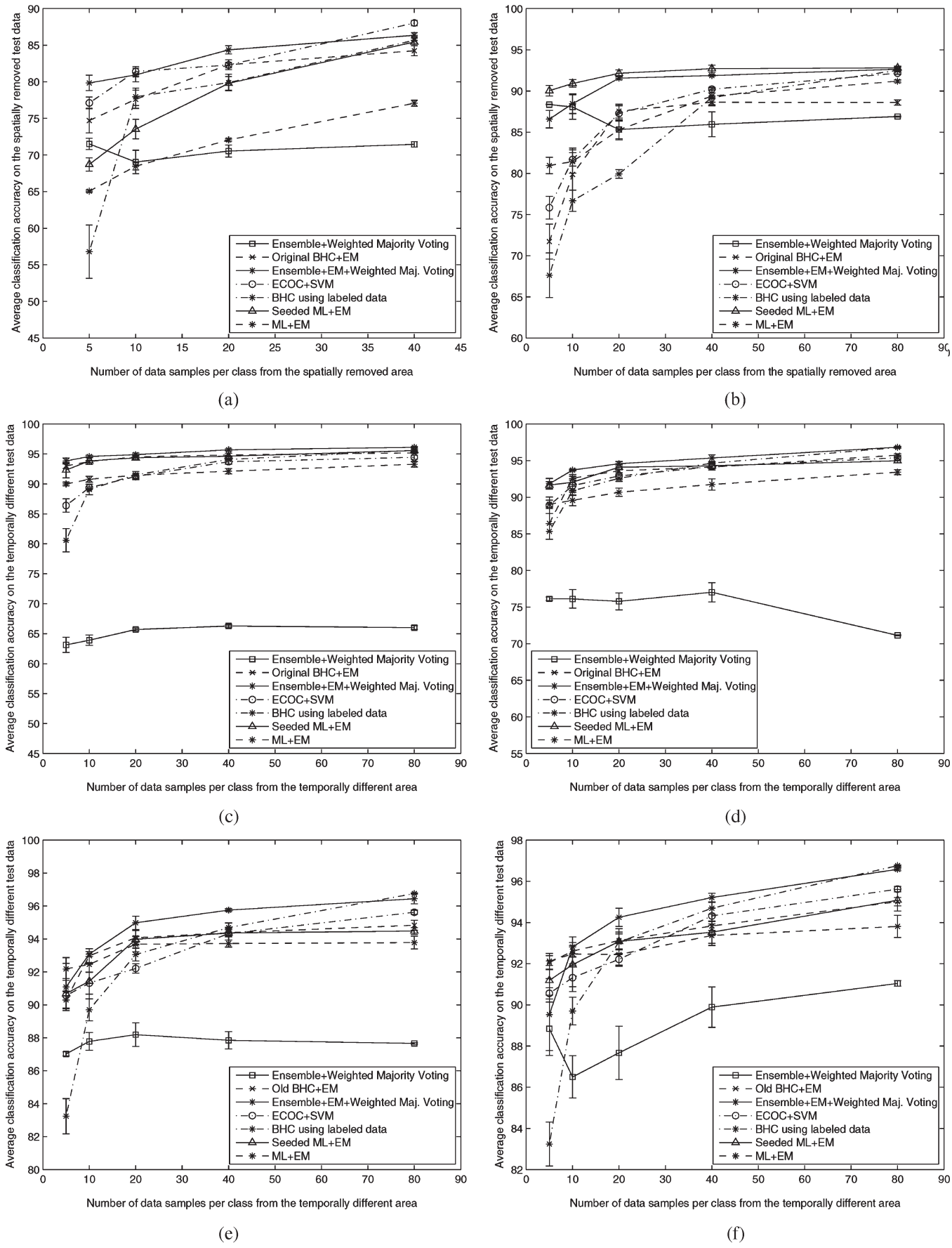

Fig. 1. Average semisupervised classification accuracies for the data sets. (a) Spatially separate KSC data. (b) Spatially separate Botswana data. (c) Botswana: May to June. (d) Botswana: May to July. (e) Botswana: June to July. (f) Botswana: May+June to July.

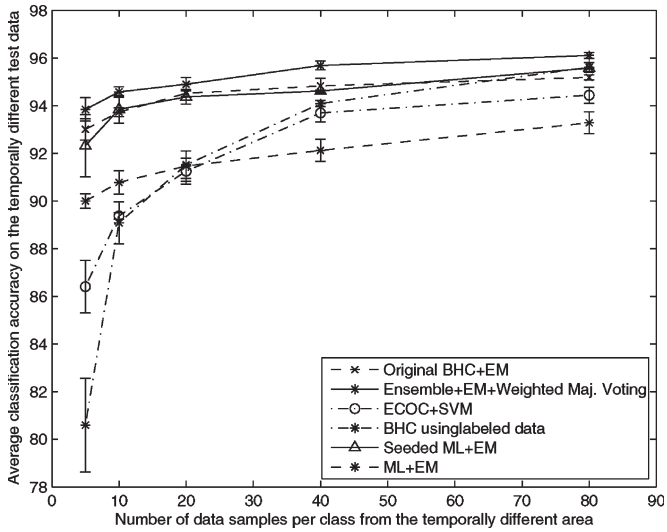

Fig. 2. Average classification accuracies for May to June data.

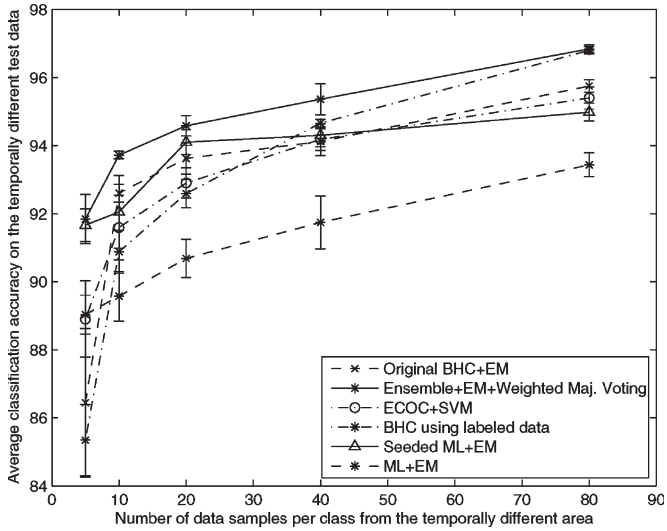

Fig. 3. Average classification accuracies for May to July data.

labeled/unlabeled data acquired from the new area. Against this combination, the alternative of building a new classifier using a powerful method (ECOC-SVM) was advantageous only when significant amounts of labeled data were available from the new areas. In addition, our approaches provide computational advantages, since fewer iterations are required for model parameters to converge because of good initialization based on prior knowledge. This study can be expanded when more hyperspectral data are available, especially to determine how the effectiveness of the knowledge transfer degrades on average, as the spatial/temporal separation of data sets is increased systematically.

ACKNOWLEDGMENT

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

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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, Jun. 2002.

[6] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. Artif. Intell. Res.*, vol. 2, no. 2, pp. 263–286, 1995.

[7] N. V. Chawla and G. Karakoulas, "Learning from labeled and unlabeled data: An empirical study across techniques and domains," *J. Artif. Intell. Res.*, vol. 23, pp. 331–336, 2005.

[8] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Mach. Learn.*, vol. 23, no. 1, pp. 69–101, Apr. 1996.

[9] L. I. Kuncheva, "Classifier ensembles for changing environments," in *Multiple Classifier Systems,* Lecture Notes in Computer Science, vol. 3077, J. Kittler and F. Roli, Eds. New York: Springer-Verlag, 2004, pp. 1–15.

[10] A. Blum, "On-line algorithms in machine learning," in *Online Algorithms: The State of the Art,* Lecture Notes in Computer Science, vol. 1442, A. Fiat and B. Woeginger, Eds. New York: Springer-Verlag, 1998.

[11] R. S. Michalski, "Toward a unified theory of learning: Multistrategy task-adaptive learning," in *Readings in Knowledge Acquisition and Learning: Automating the Construction and Improvement of Expert Systems*, B. Buchanan and D. Wilkins, Eds. San Mateo, CA: Morgan Kaufmann, 1993.

[12] K. D. Bollacker and J. Ghosh, "Effective supra-classifiers for knowledge base construction," *Pattern Recognit. Lett.*, vol. 20, no. 11–13, pp. 1347–1352, Nov. 1999.

[13] D. Silver, G. Bakir, K. Bennett, R. Caruana, M. Pontil, S. Russell, and P. Tadepalli, "Inductive transfer: 10 years later," in *Proc. NIPS Workshop*, 2005.

[14] Q. Jackson and D. A. Landgrebe, "An adpative 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.

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] ——, "Spatio temporal contextual classification of remotely sensed multispectral data," in *Proc. IEEE Int.Conf. Syst, Man, Cybern.*, Los Angeles, CA, 1990, pp. 342–344.

[23] N. Khazenie and M. M. Crawford, "Spatio-temporal random field model for contextual classification of satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 4, pp. 529–539, Jul. 1990.
[24] L. Bruzzone, R. Cossu, and D. F. Prieto, "Combining parametric and non-parametric classifiers for an unsupervised unpdating of land-cover maps," in *Multiple Classifier Systems,* Lecture Notes in Computer Science, vol. 1857, J. Kittler and F. Roli, Eds. New York: Springer-Verlag, 2000, pp. 290–299. Lecture Notes in Computer Science.
[25] 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.
[26] S. Rajan and J. Ghosh, "An empirical comparison of hierarchical vs. two-level approaches to multiclass problems," in *Multiple Classifier Systems,* Lecture Notes in Computer Science, vol. 3077, F. Roli, J. Kittler, and T. Windeatt, Eds. New York: Springer-Verlag, 2004, pp. 283–292. Lecture Notes in Computer Science.
[27] J. T. Morgan, "Adaptive hierarchical classifier with limited training data," Ph.D. dissertation, Dept. Mech. Eng., Univ. Texas, Austin, TX, 2002.
[28] S. Rajan and J. Ghosh, "Exploiting class hierachies for knowledge transfer in hyperspectral data," in *Multiple Classifier Systems*, Lecture Notes in Computer Science, vol. 3077, N. C. Oza and R. Polikar, Eds. New York: Springer-Verlag, 2005, pp. 417–428.
[29] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, May 2003.
[30] D. D. Margineantu and T. G. Dietterich, "Pruning adaptive boosting," in *Proc. 14th ICML*, Nashville, TN, 1997, pp. 211–218.
[31] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Inf. Comput.*, vol. 108, no. 2, pp. 212–261, Feb. 1994.
[32] B. Shahshahani, "Classification of multispectral data by joint supervised-unsupervised learning," Purdue Univ., West Lafayette, IN, Tech. Rep. TR-EE 94-1, 1994.
[33] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods: Support Vector Learning*, C. B. B. Scholkopf and A. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 169–184.
[34] [Online]. Available: www.lans.ece.utexas.edu/~rsuju/hyper.pdf
[35] J. T. Morgan, A. Henneguelle, J. Ham, M. M. Crawford, and J. Ghosh, "Adaptive feature spaces for land cover classification with limited ground truth data," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 18, no. 5, pp. 777–800, 2004.
[36] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," in *Proc. 17th ICML*, Stanford, CA, 2000, pp. 9–16.
[37] R. Ghani, "Using error-correcting codes for text classification," in *Proc. 17th ICML*, Stanford, CA, 2000, pp. 303–310.
[38] A. Schwaighofer. (2002). Institute for Theoretical Computer Science at Graz University of Technology. [Online]. Available: http://www.cis.tugraz.at/igi/aschwaig/software.html

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. degree in electrical engineering from the University of Texas at Austin, in 2004.

She works with the Intelligent Data Exploration and Analysis Laboratory as a Graduate Research Assistant.

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 M.S. and Ph.D. degrees from the University of Southern California, Los Angeles.

He is currently the Schlumberger Centennial Chair Professor of electrical and computer engineering with the University of Texas at Austin (UT-Austin). He joined the UT-Austin faculty in 1988. He is the Founder-Director of the Intelligent Data Exploration and Analysis Laboratory. He has authored or coauthored more than 200 refereed papers, including more than 50 full-length journal articles. His research interests are primarily 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 Program Co-Chair for the 2006 SIAM International Conference on Data Mining and the Founding Chair of the IEEE Computational Intelligence Society's Technical Committee on Data Mining. He has received ten best paper awards, including the 2005 UT-Coop Society's Best Research Paper across all departments, the Best Theory Paper at SDM 04, the Best Applications Paper at ANNIE'97, and the 1992 Darlington Award for best paper among all IEEE CAS publications.

Melba M. Crawford (M'89–SM'05) 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.

She was a Faculty Member with the University of Texas at Austin from 1990 to 2005. 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 Agriculture and Engineering. She holds the Purdue Chair of Excellence in Earth Observation. In 2004–2005, she was a Jefferson Senior Science Fellow with the U.S. Department of State. She has served as a member of the NASA Earth System Science and Applications Advisory Committee (ESSAAC) and the NASA EO-1 Science Validation team for the Advanced Land Imager and Hyperion, which received a NASA Outstanding Service Award. She also serves on the Advisory Committee to the NASA Socioeconomic Applications and Data Center, Columbia University.

Dr. Crawford is a member of the IEEE Geoscience and Remote Sensing Society, where she served as Education Director (1998–2000), Vice President for Professional Activities (1999–2001), and Vice President for Meetings and Symposia (2003–current). She is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and has been a Guest Editor for special issues on Hyperspectral Data, the Earth Observing One Mission, Advances in Methods for Analysis of Remotely Sensed Data, Landsat Missions, and Disaster Response.