# An Active Learning Approach to Knowledge Transfer for Hyperspectral Data Analysis

Suju Rajan and Joydeep Ghosh, *Fellow, IEEE*
Department of Electrical and
Computer Engineering
University of Texas at Austin
Austin, Texas 78712
Email: {suju,ghosh}@ece.utexas.edu

Melba M. Crawford, *Senior Member, IEEE*
School of Civil Engineering
Purdue University
West Lafayette, Indiana 47907
Email: mcrawford@purdue.edu

***Abstract*— Obtaining ground truth for classification of remotely sensed data is time consuming and expensive. In addition, a number of factors cause the spectral signatures of the same class to vary spatially. Therefore, successful adaptation of a classifier designed from available labeled data to classify new images acquired over other geographic locations is difficult but invaluable to the remote sensing community. In this paper we propose an active learning technique for rapidly updating existing classifiers using very few labeled data points from the new image. We also show empirically that our updated classifier exhibits better learning rates than classifiers trained via other active learning and semi-supervised methods.**

## I. Introduction

A common application of hyperspectral imaging involves mapping the spectral signatures in the images to specific land cover types. While hyperspectral data are now readily available, obtaining accurate class labels for each 'pixel' is a non-trivial task involving expensive field campaigns and time-consuming manual interpretation of imagery. Typically, 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]. 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. Hence, the naïve use of a classifier trained on available ground truth data from one region on data that are from spatially different areas, without accounting for the variability of the class signatures, will result in poor classification accuracies [2].

Existing approaches to knowledge transfer typically require labeled data from the new area for updating existing classifiers [3] [4]. A pioneering attempt at unsupervised knowledge transfer for multitemporal remote sensing images was made in [2]. In this method, classifiers trained over one image were updated using the unlabeled data from the temporally different image via Expectation-Maximization (EM) techniques [2]. The only knowledge transferred involved the estimates for the parameters of the class distributions used to initialize the EM algorithm. If the spectral signatures between the two regions vary significantly, such naïve transfer of the estimates might degrade rather than improve the EM process. A possible solution is to identify classes whose spectral signatures vary significantly between the two regions and use the corresponding labeled data to initialize those class distribution estimates. However, one then requires the labeled data from the new region to be able to identify these 'differing' classes.

In this paper, we propose an active learning approach for efficiently updating the parameters of the differing classes while using as few labeled data points from the new area as possible. The proposed method automatically identifies data points that change the current belief in the class distributions the most. Thus, labeled data are required only from those classes that vary significantly, while the existing parameter estimates are used for the remaining classes. We also empirically show that using such 'informative' data points yields better learning rates than updating the classifier with randomly chosen data points from the new area. Our proposed method is also shown to outperform other active learning strategies as well as semi-supervised EM techniques on some hyperspectral datasets.

## II. Active Learning

Let X be the set of data points, such that each data point $x_i$ has an associated class label $y_i$ drawn from a set of $k$ classes. Let us assume that there exists a pool $D$ of examples $\{x_i\}_{i=1}^{n}$ that are to be classified. In the active learning setting, the learner is provided with an initial training set drawn from $D$, say $D_L$, consisting of pairs of labeled examples, $\{x_i, y_i\}_{i=1}^{m}$. The learner then selects a $\hat{x}$ from $D_{UL} = D \backslash D_L$, such that adding $(\hat{x}, \hat{y})$ to $D_L$ and retraining the classifier minimizes the loss function associated with the classifier. 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 an $\hat{x}$ and adding it to $D_L$ is repeated for a user-specified number of iterations.

### A. Related Work

A statistical approach to active learning for function approximation problems was first proposed by Cohn et al [5], wherein the bias-variance decomposition of the squared error

0-7803-9510-7/06/$20.00 © 2006 IEEE

function was utilized in selecting $\hat{x}$. Given an unbiased learner, the goal is to select a new data point $\hat{x}$ such that adding it to $D_L$ minimizes the expected variance in the output of the learner measured over the input space. Unlike classification problems, in this case $y \in \Re^d$, and the authors present closed form solutions to compute the expected output variance using models such as the mixtures of Gaussians and locally weighted regression.

While MacKay proposed an active learning algorithm that attempts to increase the expected information gain about a user-defined variable on adding the new data point $(\hat{x}, \hat{y})$ to $D_L$ [6], the method of Roy et al [7], attempts to reduce the expected user-defined error, measured over the input space. Given a loss function $\mathcal{L}$, the true probability distribution $P_{true}(y|x)$, and the probability distribution estimated from the training set $P_{D_L}(y|x)$, the expected loss of the learner is defined as:

$$E_{D_L} = \int_x \mathcal{L}(P_{true}(y|x), P_{D_L}(y|x))P(x)dx \qquad (1)$$

where $\mathcal{L}$ is any user-specified loss function. Active learning proceeds by selecting a data point such that the expected error from using the training set $D_L^* = D_L \cup (\hat{x}, \hat{y})$ is the least, over all possible $\hat{x} \in D_{UL}$.

A popular class of active learning algorithms is that of committee based learners. Of these methods, the 'Query By Committee' (QBC) approach of [8] 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 selects the data point on which instances of the Gibbs algorithm, 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. A number of variations to the original QBC algorithm have been proposed, such as the Query by Bagging and Query by Boosting algorithm [9] and the adaptive resampling approach [10].

Active learning has also been applied in the multi-view setting [11]. In the multi-view problem, features can be partitioned into subsets each of which is sufficient for learning the mapping from the input to the output space. 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.

### B. Proposed Approach

In this paper, we propose a new active learning approach based on the method proposed by MacKay [6]. While [6] defined the 'interpolant function' as the variable whose information gain is to be maximized, we try to maximize the information gain on the posterior probability distribution defined on the set of data points. Setting $D_L^* = D_L \cup (\hat{x}, \hat{y})$ and $P^*_{D_L}(y|x)$ and $P_{D_L}(y|x)$ as the posterior probability distributions estimated from $D_L^*$ and $D_L$ respectively, it can be shown that 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 Kullback-Liebler (KL) divergence between $P^*_{D_L}(y|x)$ and $P_{D_L}(y|x)$ is maximized.

Since the true label of $\hat{x}$ is initially unknown, we follow the methodology of [5] and [7] and estimate the expected KL distance between $P^*_{D_L}(y|x)$ and $P_{D_L}(y|x)$ by first selecting an $\tilde{x} \in D_{UL}$ and assuming $\tilde{y}$ to be its label. Let $D^*_{UL} = D_{UL} \backslash \tilde{x}$ and $D_L^* = D_L \cup (\tilde{x}, \tilde{y})$. Estimating via sampling, the proposed $KL^{max}$ function can be written in terms of $(\tilde{x}, \tilde{y})$ as:

$$KL^{max}_{D_L^*}(\tilde{x}, \tilde{y}) = \frac{1}{|D^*_{UL}|} \sum_{x \in D^*_{UL}} KL(P^*_{D_L}(y|x)||P_{D_L}(y|x)) \qquad (2)$$

Note that simply assigning a wrong class label to the $\tilde{y}$ for an $\tilde{x}$ can result in a large value of the corresponding $KL^{max}_{D_L^*}$. Hence, as in [5] and [7] 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 $\hat{x}$ which maximizes this distance.

$$\hat{x} = \underset{\tilde{x} \in D_{UL}}{argmax} \sum_{\tilde{y} \in Y} KL^{max}_{D_L^*}(\tilde{x}, \tilde{y}) P_{D_L}(\tilde{y}|\tilde{x}) \qquad (3)$$

### C. Methodology

Let us assume that we have hyperspectral data from two spatially 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. To combat the effect of high dimensionality of the hyperspectral data, the feature space is reduced by recursively combining highly correlated, adjacent bands [12]. Since this 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 be exploited in Area 2.

Assuming the class-conditional density functions to be multivariate Gaussians, a Maximum Likelihood classifier is trained on the data from Area 1. Prior to learning the classifier the dimensionality of the feature space is further reduced by making use of a Fisher-m feature extractor. The best-bases feature extractor, the Fisher-M feature extractor and the ML classifier from Area 1 are then used to obtain initial posterior probabilities of the Area 2 data (E-step). While the labeled data from Area 1 are used to initialize the EM process, subsequent EM iterations are guided by the posterior probabilities assigned to the unlabeled Area 2 data. The probabilities thus obtained, are then used to update the parameters of the Gaussians (M-step). The EM iterations are performed until the average change in the posterior probabilities between two iterations is smaller than a specified threshold. A new Fisher feature extractor is also computed at each EM iteration, based on the statistics of the classes at that iteration. The updated extractor is then used to project the data into the corresponding Fisher feature space prior to the estimation of the class conditional pdfs.

Setting $P_{D_L}(y|x)$ as the posterior probability of the unlabeled data $D_{UL}$ obtained at the end of the EM iterations, we

need to select the $(\hat{x}, \hat{y})$ 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, the $(\hat{x}, \hat{y})$ 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 Eqn.2 and Eqn.3 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 the true label of $\hat{x}$.

For the next iteration of active learning, the EM process is repeated as before except that we perform constrained EM. Simply stated, in this technique while both the Area 1 and the labeled Area 2 data are used to initialize the EM algorithm, the E-step only updates the posterior probabilities for the unlabeled Area 2 data while fixing the memberships of the labeled instances according to the known class assignments.

Note that the posterior probability distributions of the Area 2 data determines $P_{D_L}(y|x)$ and guides the active learning process. Thus, we ensure that we select those '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.

## III. EXPERIMENTAL EVALUATION

In this section, we provide empirical evidence that incorporating active learning into the knowledge transfer framework results in steeper learning rate curves. We present results showing that our proposed method exhibits better learning rates than updating existing classifiers with data points selected either at random or via an existing, related active learning method. We also empirically show results that the active learning methods offer a significant advantage over the more traditional semi-supervised methods by requiring far fewer data points to obtain better classification accuracies.

### A. Data sets

The proposed active learning method was tested on hyperspectral data sets obtained from two sites: NASA's John F. Kennedy Space Center (KSC), Florida [13] and the Okavango Delta, Botswana [14].

*1) Kennedy Space Center (KSC):* The NASA AVIRIS spectrometer acquired data over the KSC on March 23, 1996. AVIRIS acquires data in 242 bands of 10nm width from 400-2500nm. The KSC data, collected from an altitude of approximately 20km, have a spatial resolution of 18m. Removal of noisy and water absorption bands resulted in 176 candidate features. 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 test set (Area 2) is a different subset of the flight line than the $512 \times 614$ data set from Area 1. While the number of classes in the two regions differs, we restrict ourselves to those classes that are present in both regions.

*2) Botswana:* This $1476 \times 256$ pixel study area is located in the Okavango Delta, Botswana, and has 14 different land cover types. Data from this region were obtained by the NASA EO-1 satellite for the calibration/validation portion of the mission in 2001. The Hyperion sensor on EO-1 acquires data at 30m pixel resolution over a 7.7km strip in 242 bands covering the 400-2500nm portion of the spectrum in 10nm windows. Uncalibrated and noisy bands that cover water absorption features were removed resulting in 145 features. 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.

### 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 separate region (Area 2).

The ML-EM classifier, for all the methods evaluated, was modeled using a multivariate Gaussian for each class. 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. Constrained EM was used to update the parameters of the Gaussians as well as the Fisher discriminant as detailed in Section II-C.

The proposed active learning method was evaluated against the baseline method of choosing the data points, one at a time, at random from Area 2 and using constrained EM to update the estimates of the class parameters from Area 1.

The active learning approach of Roy et al. [7] was also implemented and evaluated. This method attempts to reduce the expected error measured over the input space. Using the log-loss function results in selecting those data points that cause an increase in the future expected entropy. Following the notation from Eqns. 2 and 3, the $\hat{x}$ is selected using the following equations:

$$E_{D^*_L}(\tilde{x}, \tilde{y}) = \frac{1}{|D^*_{UL}|} \sum_{x \in D^*_{UL}} \sum_{y \in Y} P_{D^*_L}(y|x) log P_{D^*_L}(y|x) \quad (4)$$

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

$$\hat{x} = \underset{\tilde{x} \in D_{UL}}{argmin} \sum_{\tilde{y} \in Y} E_{D^*_L}(\tilde{x}, \tilde{y}) P_{D_L}(\tilde{y}|\tilde{x}) \quad (5)$$

Finally, for the semi-supervised scenario, small quantities of labeled data were selected from each class at random. The knowledge transfer method as proposed in [2] was modified to incorporate the labeled data into the EM process. The Area 1 data and the labeled Area 2 data were used to initialize the Gaussians prior to performing the EM iterations. The parameters of the Gaussians and the Fisher feature extractors

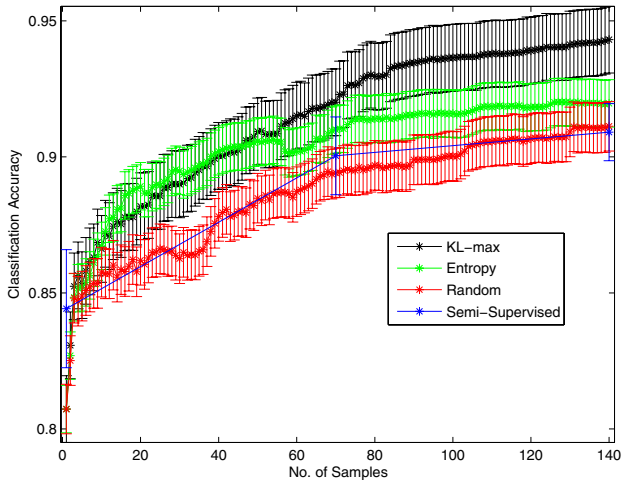


Fig. 1. Learning Rates for Botswana Area 2

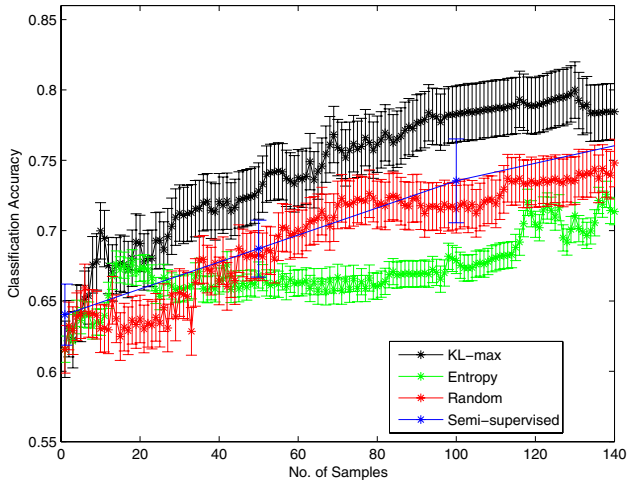


Fig. 2. Learning Rates for KSC Area 2

were then updated using the labeled and unlabeled data from Area 2 via constrained EM.

## IV. Results and Discussion

Figure 1 and 2 show the learning rate curves for the Botswana and KSC datasets over 140 active learning iterations. It can be clearly seen that in both cases the proposed active learning approach yields better classification accuracies than the 'entropy' method of Roy et al. [7]. It is interesting to note that adding a single randomly chosen data point at a time (Random curve) has the same effect as batch-training the classifier via the semi-supervised technique (Semi-Supervised curve). Figure 2 shows that for the KSC dataset the entropy-based approach performs worse than the random active learning method, this is because there is a greater disparity in the spectral signatures of the classes between the two areas. The proposed method, however, remains unaffected by the magnitude of spatial separation in the datasets.

## V. Conclusion

We proposed a new active learning based knowledge transfer approach which seems to be particularly well-suited to the scenario in which the distributions of the classes show spatial (or temporal) variations. The principle of selecting those data points that change the existing belief in class distributions the most helps in efficiently and rapidly updating the existing classifier for a new, related problem. The proposed method is empirically shown to be far better than choosing random points, batch semi-supervised methods, and an entropy-based active learning method. This study 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 increased systematically.

**Acknowledgment:** This work was supported by NSF Grant IIS-0312471.

## 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. and Remote Sens.*, vol. 32, no. 5, pp. 1087–1095, 1994.

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

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

[4] 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. 2004 Intl. Geosci. and Remote Sens. Symposium (IGARSS)*, Anchorage, USA.

[5] D. Cohn, Z. Gharamani, and M. Jordan, "Active learning with statistical models," *Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

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

[7] N. Roy and A. K. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proc. 18th Intl. Conf. on Machine Learning (ICML)*, Williams Collge, USA, 2001, pp. 441–448.

[8] H. S. Seung, M. Opper, and H. Smopolinsky, "Query by committee," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, Pittsburgh, USA, 1992, pp. 287–294.

[9] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proc. 15th Intl. Conf. on Machine Learning (ICML)*, 1998, pp. 1–9.

[10] V. S. Iyengar, C. Apte, and T. Zhang, "Active learning using adaptive resampling," in *Proc. 6th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*. Boston, USA: ACM Press, 2000, pp. 92–98.

[11] I. Muslea, S. Minton, and C. Knoblock, "Active + Semi-supervised learning = Robust muti-view learning," in *Proc. 19th Intl. Conf. on Machine Learning (ICML)*, Sydney, Australia, 2002, pp. 435–442.

[12] S. Kumar, J. Ghosh, and M. M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. Geosci. and Remote Sens.*, vol. 39, no. 7, pp. 1368–79, 2001.

[13] J. T. Morgan, "Adaptive hierarchical classifier with limited training data," Ph.D. dissertation, Dept. of Mech. Eng., Univ. of Texas at Austin, 2002.

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