AN EFFICIENT ACTIVE LEARNING ALGORITHM WITH KNOWLEDGE TRANSFER FOR HYPERSPECTRAL DATA ANALYSIS

Goo Jun *Joydeep Ghosh*

Department of Electrical and Computer Engineering
The University of Texas at Austin, Austin TX 78712, USA
{gjun, ghosh}@ece.utexas.edu

ABSTRACT

We propose an active learning algorithm with knowledge transfer for classification of hyperspectral remote sensing data. The proposed method is based on a previously proposed algorithm, but yields faster learning curves by adjusting distributions of labeled data differently for the old and the new data. With the proposed method, the classifier can effectively transfer its knowledge learned from one region to a spatially or temporally separated region whose spectral signature is different. Empirical evaluation of the proposed algorithm is performed for two different hyperspectal datasets.

Index Terms— classification, hyperspectral data, active learning, knowledge transfer

1. INTRODUCTION

Training a classifier for characterizing land cover based on hyperpectral imagery usually requires large amounts of labeled data. Obtaining ground truth class labels of a remote sensing image is expensive. Moreover, there are temporal and spatial variations in the spectral signatures due to many reasons such as seasonal effects, ecological or topographical variations, weather conditions, and geological differences. Since it is impractical to obtain ground truth of all areas at multiple times, we need "transfer learning" techniques that can achieve high classification accuracy with relatively small number of labeled samples from a new area by exploiting previously processed information [1].

Active learning is a method of online learning, where a learner strategically selects new training examples that provide maximal information about the unlabeled dataset, resulting in higher classification accuracy for a given training set size as compared to using randomly selected examples. Active learning is most useful when there are sufficient number of unlabeled samples but it is expensive to obtain class labels. Most active learning algorithms however assume that the model built upon labeled data is not biased, and the probability distribution of the unlabeled and existing datasets are identical. These assumptions do not hold for remote sensing applications under spatial and temporal variations; hence we need to incorporate transfer learning techniques into an active learner. Rajan *et al.* [1] recently proposed KL-max algorithm to transfer knowledge with active learning for hyperspectral data, setting the current state of the art. In this paper, we build on Rajan *et al.*'s approach for more effective knowledge transfer using active learning.

This work was supported by NSF Grant IIS-0705815.

2. ACTIVE LEARNING

Having enough number of labeled examples is important to obtain a good classifier, especially for difficult problems. In many cases, however, acquiring ground truth for large number of examples is an expensive and time-consuming task. On the contrary, unlabeled samples are easier to obtain for some problems. For example, classification of web pages belongs in the category. A simple web crawling robot can automatically collect huge amount of web pages, or unlabeled samples, without much difficulty or cost involved. Labeling all of the collected web pages, however, requires a lot of effort, and it is virtually impossible for a single human expert. For land cover classification based on remotely sensed data, a similar situation is encountered. Airborne or satellite images usually cover large geographical areas, while finding the actual land cover type is costly and involves efforts of human experts.

In the active learning literature, conventional learning algorithms without active selection is often referred as 'passive learning' in contrast to active learning algorithms. In passive learning, a training set is usually selected randomly from the entire data. In active learning, a learner chooses k examples those are considered most useful, obtains ground truth for them, learns from these k examples, and then repeats the choose-and-learn process. Query-by-committee (QBC) [2] is a well-known active learning algorithm that employs a committee of independent classifiers, and it is shown that the algorithm guarantees positive information gain for each query under several assumptions, while the information gain from randomly selected examples converges to zero asymptotically. MacKay [3] proposed an active learning framework, where

the learner chooses an example which has the most expected informational gain. Lewis and Gale [4] proposed a sampling criteria for the active learning, called uncertainty sampling, and various kinds of uncertainty measures can be used depending on the problem domain. Cohn *et al* [5] proposed a method based on a statistical analysis of the active learning problem, where the point that minimizes the variance of a model is selected to be labeled. In general, most active learning algorithms aim to achieve lower error rate than passive learning with same or fewer number of labeled samples.

3. KL-MAX

Classification of land cover types with remotely sensed hyperspectral imagery mostly depends on the spectral signature of each land cover type, which has temporal and spatial variations as discussed in section 1. It is not practical to build a new classifier whenever temporal or spatial change occurs, because training a new classifier requires large amounts of labeled data, even with active learning. A positive aspect of the problem is that spectral signatures of a new region are not completely different from those of the old region when spatial or temporal difference is small. If we could effectively reuse our knowledge derived from previous data, then a classifier for the new area can be trained with significantly less number of samples. Applying a previously trained classifier directly to the new region, however, often results in poor classification accuracies and it also degrades performances of active learning algorithms. Most active learning algorithms require an initial model, and it is assumed that the initial model is built upon a distribution identical to that of unlabeled data. For example, Cohn et al's approach requires that the model is not biased [5], and MacKay's approach is also based on the correctness of the initial model [3]. For this reason, we need to adapt our model for the new region while maintaining its useful knowledge by employing transfer learning techniques.

In our setup, it is assumed that there exist two different datasets from temporally or spatially distant regions, that we refer to as areas 1 and 2 respectively. We denote our set of labeled samples from area 1 as D_L, and we have a model trained on D_L, which is used as an initial model in the subsequent active learning process to select a sample from D_{UL}, a set of unlabeled data from area 2. The difficulty of this approach arises when the probabilistic distribution of D_L is different from that of D_{UL}. If our model built upon the labeled set does not provide unbiased results on the new set, then we cannot expect samples selected from the new set using traditional active learning to be the most informative samples, which results in a slower learning curve. If we could build a better model by using D_L and D_{UL} together, then we could choose more informative samples. Having more informative samples leads the model to be more accurate on D_{UL}, consequently enabling better choice of unlabeled samples again, and it forms a positive feedback for a faster learning curve. In this manner, the KL-max algorithm [1] effectively combines the active learning strategy with transfer learning.

The KL-max algorithm transfers knowledge in a semi-supervised manner. The class-conditional distribution of D_L is assumed to be multi-variate Gaussian, and is estimated by maximum likelihood (ML). The estimated distribution is then used to initialize expectation-maximization (EM) process on the unlabeled data to obtain a posterior probability distribution of the unlabeled dataset, $P_{D_L}(y|x)$. The active learning algorithm used in KL-max is based on MacKay's approach [3], and it selects a data point $(\hat{x}, \hat{y})$ that maximizes the information gain on the posterior probability distribution. The information gain between two posterior distributions $P_{D_L^*}(y|x)$ and $P_{D_L}(y|x)$ can be measured by the Kullback-Liebler (KL) divergence between $P_{D_L{}^*}(y|x)$ and $P_{D_L}(y|x)$. Because we do not the true label $\hat{y}$ for $\hat{x}$, the expected KL divergence is calculated over all possible class labels $\tilde{y} \in Y$.

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

Defining $D_{UL^*} = D_{UL} \setminus x^*$ and $D_L{}^* = D_L \cup (\tilde{x}, \tilde{y})$, the KL^{max} can be written in terms of $(\tilde{x}, \tilde{y})$ as:

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

After obtaining the new data point $(\hat{x}, \hat{y})$, the ML-EM process is repeated with the augmented labeled dataset, followed by constrained EM iterations. KL-max algorithm shows faster learning rates than several other active learning algorithms for hyperspectral data [1].

4. PROPOSED METHODOLOGY

The performance of the KL-max algorithm can be greatly improved if we provide more accurate initial distribution for the EM process in the ML-EM framework. In the KL-max algorithm, all samples from area 1 and new samples from area 2 are treated equally for ML estimation, although their distributions could be significantly different from each other. As a result, the estimated distribution is much closer to the distribution of area 1, since we have only a small fraction of samples from area 2 compared to the number of samples from area 1.

Recently, a boosting algorithm for transfer learning, TrAdaBoost, was proposed by Dai *et al* [6]. TrAdaBoost is a transfer learning method based on the AdaBoost algorithm, where more weights are given to samples misclassified by a base learner and another base learner is subsequently trained under the modified distribution to form an ensemble of base classifiers. TrAdaBoost does not equally increase weights of all samples misclassified by a base learner, but increases weights of misclassified samples belonging to the new dataset, and decreases weights of misclassified samples

belonging to the old dataset. In this paper, we propose a method based on the same philosophy as in [6], modified for an online active learning environment.

In active learning, we obtain an updated classifier whenever a new labeled sample is acquired. The updated classifier is assumed to be more trustworthy than previous ones, since it is trained with more information. Consequently, a cumulative update as in boosting is not appropriate, since new weights are largely affected by previous weights obtained from less accurate classifiers. Some samples initially thought to be bad could turn out to be useful in later stages as we gather more information on the new distribution, and vice versa. Therefore, we construct a new distribution of weights for each classifier, instead of applying cumulative updates.

Many different ways are possible for weight distribution. The baseline strategy is to assign lower weights for misclassified samples in D_L, and higher weights for misclassified samples in D_N, the set of newly acquired labeled samples. The proposed method is based on some qualitative analyses, depending on the number of new and old labeled samples. Our first analysis is based on the assumption that having more samples in D_N results in a more reliable classifier, which makes it more convincing that misclassified samples from D_L are less useful. Therefore, lower weights should be assigned to misclassified samples in D_L as we get more samples in D_N. Another observation is that although emphasizing misclassified points in D_N accelerates the transfer learning process initially, eventually it can make the classifier sensitive to outliers or overfitted. For that reason, after we get enough number of samples in D_N, we should gradually decrease weights of misclassified samples in D_N.

In the proposed methodology, the weight updating rules were determined heuristically after exploring several algorithms based on the aforementioned qualitative observations. Note that D_L^* is an augmented set of labeled data, $D_L^* = D_L \cup D_N$. Suppose w_i is the weight associated with data point x_i, where $(x_i, y_i) \in D_L^*$ and $h^* : X \rightarrow Y$ is the current hypothesis. Weights for sample points in D_L^* are calculated as:

1) if $(x_i, y_i) \in D_L$ and $h^*(x_i) \neq y_i$,

$$w_i = (1 + \log |D_N|)^{-1}$$

2) if $(x_i, y_i) \in D_N$ and $h^*(x_i) \neq y_i$,

$$w_i = 1 + \frac{\epsilon_N}{1 - \epsilon_N} \cdot \log \left[|D_N| \cdot (|D_L| - |D_N|) \right] \mathbf{1}_{(|D_N| < |D_L|)}$$

3) if $h^*(x_i) = y_i$, $w_i = 1$.

$\mathbf{1}_{(|D_N| < |D_L|)}$ is an indicator function, making $w_i = 1$ for $(x_i, y_i) \in D_N$ when $|D_N| \geq |D_L|$. The parameter ϵ_N is the error rate measured on the set D_N, and the weight also gets close to 1 when ϵ_N is very small. The mean and the covariance of the class-conditional distribution, $P_{D_L}(x|y)$ are estimated using weighted ML.

5. EXPERIMENTS

5.1. Data

The proposed method was evaluated on hyperspectral datasets taken from NASA's John F. Kennedy Space Center(KSC), Florida [7] and the Okavango Delta, Botswana [8].

1) Kennedy Space Center (KSC): The NASA Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) data originally consists of 242 bands , but only remaining 176 bands are used after removing noisy and water absorption bands. There are 13 different land cover types including water and mixed classes, which causes more difficulties in the classification. Two different subsets of flight line, each with 512×614 pixels with 18m spatial resolution, are used for experiments.

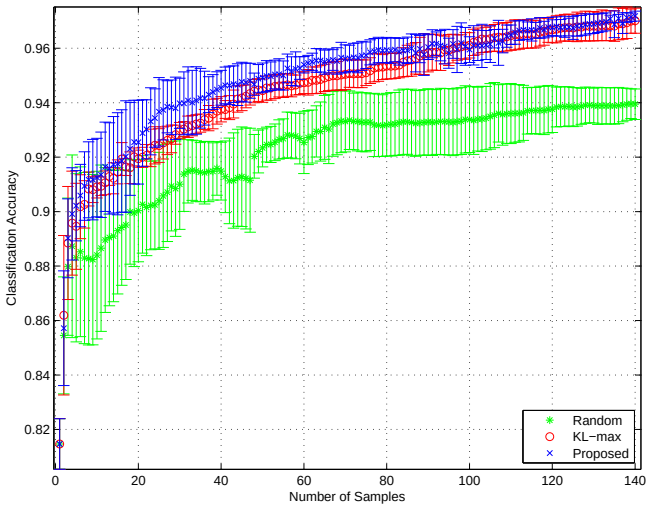
2) Botswana: Data from the Okavango Delta were obtained by the NASA EO-1 satellite with the Hyperion sensor on May 31, 2001. The area used for experiments has 1476×256 pixels with 30m spatial resolution, with 14 different land cover classes. The acquired data originally have 242 bands, and pre-processing of data resulted in 145 remaining bands. Two spatially disjoint datasets are sampled from the original hyperspectral image, and used as area 1 and 2 data.

5.2. Experimental Methods

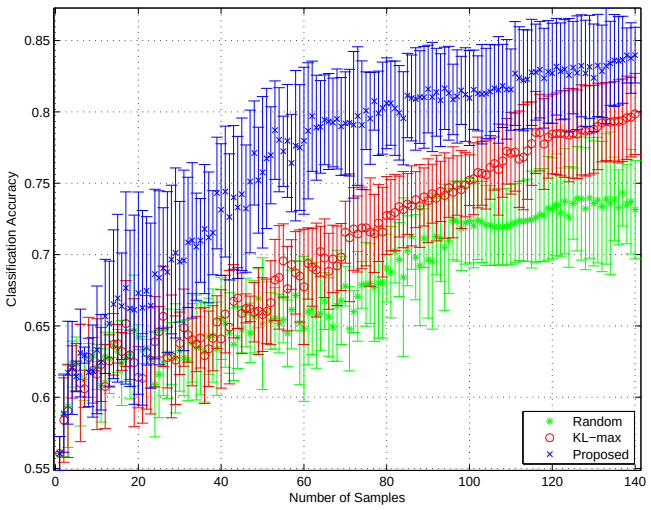
Both datasets include hyperspectral data from two distant areas in the region. Each dataset from area 1 is randomly sampled at 75% to construct a training set, D_L and remaining 25% were used as validation data. Samples from area 2 were used as unlabeled dataset, D_{UL}. Best-bases feature extraction [9] and Fisher's linear discriminant analysis are used for dimensionality reduction. The number of bases for the best-bases feature extraction was determined by using the validation set of area 1, and the number of bases with highest classification accuracy was selected. Roy and Maccullum's sampling method [10] was employed to reduce number of samples used in the active learning. Class-conditional distribution was assumed to be multi-variate Gaussians, and was estimated on the area 1 dataset using ML estimation. Posterior distribution, $P_{D_L}(y|x)$ was obtained after EM iterations, and a new labeled sample $(\hat{x}, \hat{y})$ is obtained as described in section 3. With the new labeled data, distribution of the labeled samples are re-weighted as in section 4. The re-weighted distribution is then used as the initial distribution of subsequent constrained EM process. Each experiment is repeated for 10 times to obtain average accuracies and standard deviations.

6. RESULTS

Fig. 1 shows average classification accuracies and standard deviations from the proposed algorithm, KL-max, and the baseline method, where samples from D_{UL} are randomly (passively) picked. In Fig. 1-(a), learning rates of



(a) Botswana Learning Curves



(b) KSC Learning Curves

Fig. 1. Experimental Results for Botswana and KSC Data

the proposed algorithm do not show significant improvement from the KL-max algorithm for Botswana dataset while both curves are far better than the baseline method. In the KSC result, the gap between the proposed method and KL-max is larger than the gap of KL-max and the baseline (random) method. Figure 1-(b) shows that for the KSC dataset the proposed approach performs much better than the KL-max active learning method. This is because there is a greater disparity in the spectral signatures of the classes between the two areas of KSC data than areas of Botswana data.

7. CONCLUSION

In this paper, we proposed an algorithm for efficient active learning with transferred knowledge based on the KL-max algorithm by adjusting distributions of the labeled dataset. The proposed method provides substantially superior empirical results when the discrepancy between the labeled and unlabeled dataset is significant.

8. REFERENCES

[1] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, 2008.

[2] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, PA, USA, 1992, pp. 287–294, ACM Press.

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

[4] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1994, pp. 3–12, Springer-Verlag New York, Inc.

[5] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH*, vol. 4, pp. 129, 1996.

[6] W. Dai, Q. Yang, G. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th International Conference on Machine Learning*, Corvalis, Oregon, 2007, pp. 193–200, ACM.

[7] J. T. Morgan, *Adaptive hierarchical classifier with limited training data*, Ph.D. thesis, Univ. of Texas at Austin, 2002.

[8] 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.

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

[10] Nicholas Roy and Andrew McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proceedings of the 18th International Conference on Machine Learning*. 2001, pp. 441–448, Morgan Kaufmann Publishers Inc.