Exploiting Class Hierarchies for Knowledge Transfer in Hyperspectral Data

Suju Rajan and Joydeep Ghosh

Dept. of Electrical and Computer Engineering, University of Texas at Austin, Austin
TX 78712, USA.
{rsuju,ghosh}@lans.ece.utexas.edu

Abstract. Obtaining ground truth for hyperspectral data is an expensive task. In addition, a number of factors cause the spectral signatures of the same class to vary with location and/or time. Therefore, adapting a classifier designed from available labeled data to classify new hyperspectral images is difficult, but invaluable to the remote sensing community. In this paper, we use the Binary Hierarchical Classifier to propose a knowledge transfer framework that leverages the information gathered from existing labeled data to classify the data obtained from a spatially separate test area. Experimental results show that in the absence of any labeled data in the new area, our approach is better than a direct application of the old classifier on the new data. Moreover, when small amounts of labeled data are available from the new area, our framework offers further improvements through semi-supervised learning mechanisms.

1 Introduction

The deployment of hyperspectral sensors on-board the earth observing satellites has generated large amounts of remotely sensed data, providing detailed hyperspectral images over extensive regions of the earth. A typical application is to determine the land cover types corresponding to the spectral signatures in the hyperspectral image, which can then be used, for example, to monitor changes in the ecosystem over large geographic areas. However, while large quantities of hyperspectral data are now available, obtaining reliable and accurate class labels for each 'pixel' is a non-trivial task, involving either expensive field campaigns or time-consuming manual interpretation of the hyperspectral images. Currently, researchers obtain the class labels for a few pixels in an image (which typically have 100,000+ pixels) and then attempt to label some other pixels in that same image. Characterizing a new image is treated as a separate, independent classification problem. Note that factors such as atmospheric and light conditions, topographic variations, etc., alter the spectral signatures corresponding to the same land cover type over images acquired at different times and/or different regions. Hence, naïve use of a classifier trained on data from one area to either spatially or temporally different data without accounting for the variability of the class signatures results in poor classification accuracies [2] [4]. Theoretically, an ideal approach would be to pool all images of interest and then extract training data sampled uniformly at random from this pool to form a classifier that works on all these images. But a host of real-life issues such as non-availability of all the data at a given time, ownership, poor performance due to spatial variability of class signatures, size of data, etc., currently prevent researchers from being able to follow this path.

In this paper, we study a more feasible middle ground of how to exploit the knowledge inherent in a classifier trained over one area to help classify the data obtained (perhaps at a later time) from a spatially separate area. Our framework exploits additional knowledge from existing classifiers. Specifically, we use a multi-classifier system called Binary Hierarchical Classifier (BHC) [8] for this purpose. The BHC automatically derives a hierarchy of the target classes based on their mutual affinities. This hierarchy, along with the knowledge of the features extracted at each node of the BHC tree, can be used to facilitate the unsupervised classification of new data from the *spatially separate* area. Besides the unsupervised setting, the framework presented can also be employed in a semi-supervised scenario when very small quantities of labeled data are available from the spatially separate area. We present results of experiments that demonstrate the advantages of our proposed framework over other powerful multi-classifier systems, such as the ECOC [3], for the purposes of knowledge transfer for hyperspectral data.

2 Related Work

Given that obtaining labeled data, especially for remote-sensing applications is difficult and time-consuming especially for remote areas, any scheme that can leverage existing labeled data even from an area other than the one under consideration, is very desirable. The advantages of using unlabeled data in a semi-supervised setting to mitigate the sample size problems for hyperspectral data was first studied in [16]. In this work both the labeled and the unlabeled data samples came from the same image, i.e. from a common underlying distribution.

A related problem was addressed in the context of temporally varying remote sensing images in [2]. Given an image t_1 of a certain land area with a labeled training set, the problem was to classify another image t_2 of the same land area obtained at a different time. A maximum likelihood classifier was first trained on the labeled data from t_1, assuming normal distribution of the class-conditional density functions. The mean vector and the covariance matrix of the classes from t_1 were used as initial approximations to the parameter values of the same classes from t_2. These initial estimates to the classes from t_2 were then improved via EM using the corresponding unlabeled data. Experimental results on a couple of multispectral images acquired at different times showed that utilizing the labeled data from t_1 yielded comparable classification accuracy with that of a maximum likelihood classifier trained on the labeled data from t_2.

Several online-learning algorithms [1] have also been proposed to deal with the problem of temporally varying data distributions. Various possibilities of applying such online techniques in a multi-classifier setting are outlined in [9]. Most approaches to the problem of population drift design the classifier as a feedback system, in which it is assumed that there is a steady stream of objects whose true labels are revealed immediately after classification by the existing classifier. This additional knowledge of the true class labels may then cause a change in the existing classifier [6]. While hyperspectral data obtained over extensive regions (or different times) also faces a similar problem of 'population drift', unlike the on-line frameworks, one does not have access to a streaming set of labeled data samples.

While [2] demonstrates the advantage of using previously acquired knowledge in classifying a novel image, the amount of knowledge transferred was restricted by the classifier under consideration, namely the maximum likelihood classifier. The only knowledge from the training data that was transferred in that framework, were the estimates of the distributions of the class density functions in the original feature space. 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.

2.1 Binary Hierarchical Classifier

The BHC [8] is a multi-classifier system that was developed primarily to deal with multi-class hyperspectral data. The BHC involves recursively decomposing a multi-class (C-classes) problem into (C-1) binary meta-class problems, resulting in (C-1) classifiers arranged as a binary tree. The given set of classes is first partitioned into two disjoint meta-classes, and each meta-class 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 two-meta-classes is not arbitrary, but is obtained through a deterministic annealing process, which encourages similar classes to remain in the same partition. Thus, as a direct consequence of the BHC algorithm, classes that are similar in the input feature space are lumped into the same meta-class higher in the tree. Interested readers are referred to [8] for details of the algorithm. Each internal node of the BHC utilizes a Fisher discriminant and a Bayesian classifier. To combat the small sample size problems, the dimensionality of the feature space is reduced by recursively combining highly, correlated adjacent hyperspectral bands [7]. This best-bases method of feature extraction makes use of class information, as the correlation between bands varies among the classes, thereby yielding an interpretable feature space.

Recent empirical evaluations have shown that the BHC offers comparable classification accuracies with that of other multi-classifier systems such as the ECOC [15]. Moreover, the BHC also reveals a lot of knowledge inherent in the training data. The hierarchy of classes, for instance, might be useful as the relationships between classes in one area might still hold in another new area. Further, since the best-bases feature extraction method makes use of class-specific information in deciding 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.

3 Knowledge Transfer Framework

Let us assume that we have hyperspectral data from two spatially separate 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 is 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 (semi-supervised case).

3.1 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.

The first approach was to use the hierarchy of classes and the best bases feature extractors of the area 1 solution, but modify the binary classifiers in this multiclassifier system to account for the changed statistics of the spatially separate data. This was achieved via the EM framework [14] in which the training data was used to initialize the EM algorithm and the spatially separate data from area 2 was treated as the unlabeled data. Mixtures of Gaussians were used at each node of the BHC tree, with the number of Gaussians corresponding to the number of classes at each node. Each Gaussian was initialized with the mean and the covariance of the corresponding class in the training data. The initial estimates were then used to determine the posterior probabilities of the corresponding classes in the spatially separate area. EM iterations were performed until the average change in the posteriors between two iterations fell below a threshold [2] [14]. Thus, an updated Fisher-m feature extractor was computed at each node based on the statistics of the meta-classes at that iteration.

An analysis of the results showed that, while this approach was somewhat better than a direct application of the old classifier, the errors were mostly concentrated in a few classes. A closer inspection revealed that the spectral signatures of these classes had changed sufficiently enough 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 so that some of them would be more suitable for the new area.

Thus our second approach was to introduce different randomizations into the training data and generate a BHC tree for each such randomization. The design space for the BHC tree also offers a lot of possibilities for randomizing its tree generation process. Some of the factors that were varied are the percentage of available training data, the number of features available at each node and also the process of generating meta-classes (top-down versus bottom-up) [8]. To account for the possibility of changing priors of classes, another set of BHC trees was generated by randomly altering the priors of the classes in the training data. Bagging - a popular method for generating classifier ensembles was also used to generate a set of BHC trees. Finally, in an attempt to account for the changes in class spectral signatures, a fourth set of classifiers was generated by randomly switching the labels of a small percentage of the data-points for each class. Note that in the absence of any labeled data from area 2, there is no way of evaluating which of the randomly generated BHC trees best suits the spatially separate 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 Q-diversity measure [10], which indicates the degree of correlation between a pair of classifiers, was used to ensure the diversity of our classifier ensemble. Each tree in the classifier ensemble was made to label the data from area 2 and these labels were then used to obtain the Q- diversity measure between each pair of classifiers. The classification results of a smaller set of classifiers with the lowest average pairwise Q-measure (i.e., higher diversity) were then combined via simple majority voting.

3.2 Semi-supervised Case

If there are adequate amounts of labeled data from area 2, one can just train a classifier using the available labeled data. But for small amounts of labeled data, we would expect the two knowledge transfer mechanisms discussed earlier to be superior especially if they exploit this added information. In this section, we generalize both knowledge transfer methods in order to leverage the labeled data, and also determine how much labeled data is required from the spatially separate area before the advantages of transferring information from the old solution disappear.

The EM-based method was modified to perform constrained EM instead. Simply stated, in this technique, 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.

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 Q-diversity measure, we went through another round of pruning to include only those classifiers with higher classification accuracies on the labeled data. A scheme similar to the on-line weighted majority algorithm as detailed in [11] was used to weight the different classifiers. In the weighted majority algorithm, all classifiers are assigned a weight. Prior to learning, the weights of all the classifiers are the same, and then as each data sample is presented to the ensemble, a classifier's weight is reduced if it misclassifies that example. At the end of this learning, all those classifiers that have better classification accuracies on the incoming data will have higher weights. Then for each new example, the ensemble returns the class that gets the maximum total weighted vote over all the classifiers. This weighted majority 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 [11].

The labeled data was also used to initialize the mean vectors and the covariance matrices of the meta-classes, at the nodes of the binary trees in the Q-diversity measure pruned ensemble. The labeled and the unlabeled data from area 2 were then used for constrained EM in each of the binary trees. The classification results of the resulting ensemble were then combined using the weighted majority algorithm as detailed above.

4 Experimental evaluation

In this section, we provide empirical evidence that in the absence of labeled data from the spatially 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 labeled data from the new area, our framework performs better than the current state-of-the-art ECOC multi-classifier system [3] with SVMs [5] as the binary classifiers.

4.1 Datasets

The knowledge transfer approaches described above were tested on hyperspectral datasets obtained from two sites: NASA's John F. Kennedy Space Center (KSC), Florida [12] and the Okavango Delta, Botswana [4]. In both these datasets, the labeled data (area 1) was subsampled such that 75% of the data was used for training and 25% as the test set. For both cases, a second test set was also acquired from a spatially separate region (area 2). Since the spatially separate test set comes from a different geographic location, various factors such as the sun angle, shadow and other temporal factors cause a natural variation of the hyperspectral signatures. This variation in spectral signatures along with the changes in the apriori probabilities of the landcover classes offers an ideal setting to test the knowledge transfer framework. While the numbers of classes in the two regions vary, we restrict ourselves to those classes that are present in both regions.

4.2 Experimental Methodology

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 [7] prior to applying the Fisher feature extractor.

	Baselines			Knowledge Transfer Approaches	
Name	Old BHC	Old ECOC +SVM	Ensemble BHC +Maj. Vote	Old BHC + EM	Ensemble BHC +EM +Maj. Vote
KSC	61.84 (0.60)	64.27 (0.27)	64.4 (0.10)	65.82 (2.86)	65.12(1.97)
Botswana	73.04 (2.25)	75.22(0.65)	74.82(0.75)	79.13 (1.96)	79.8 (1.8)

Table 1. Average unsupervised classification accuracies for the spatially separate test sets .

Adjacent bands were merged until the ratio of the training samples to the number of dimensions was at least five at each node of the classifier tree [13]. For both the unsupervised and the semi-supervised cases, the classification accuracies were obtained by averaging over 5 different samples of the training data (from area 1) or the labeled spatially separate data (from area 2) as the case may be.

The ensemble of BHC trees was generated by varying the percentages of available training data (5 different rates), the number of features available at each node (10 different subsets), the top-down and bottom-up generation of the BHC tree, by randomly altering the priors of the classes in the training data (40 such randomizations), bagging (40 samplings with replacement) and by randomly switching the labels of a small percentage (10%) of the data-points for each class (40 such randomizations). Thus, a total of 220 different randomizations of the BHC were generated from the original training data. The Q-diversity measure was then used to prune the existing ensemble such that the final ensemble contained the 10 classifiers with the lowest average pairwise Q-measure.

For the ECOC system, the code matrix was generated using the technique proposed in [3]. SVMs with Gaussian kernels were trained for each binary problem induced by the code matrix. The The implementation and the tuning of the SVM classifiers followed the same method as in [15] with 40% of the available labeled data as the validation data.

4.3 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 the spatially separate area 2. Table 1 shows the classification accuracies obtained by the baseline and the knowledge transfer approaches on the area 2 data. As a frame of reference, the classification accuracies on the area 1 test set for the BHC and the ECOC-SVM are 93.05%(±1.17) and 93%(±1.03) for the KSC dataset. For the Botswana dataset the corresponding classification accuracies are and 94.52%(±0.79) and 95.63%(±0.95) respectively.

It can be seen from Table 1 that when the unlabeled data from area 2 is used via EM to update the statistics of the meta-classes in the BHC tree, the resulting BHC tree performs better than the old BHC. However, updating the BHC ensemble via semi-supervised learning does not provide any additional gains.

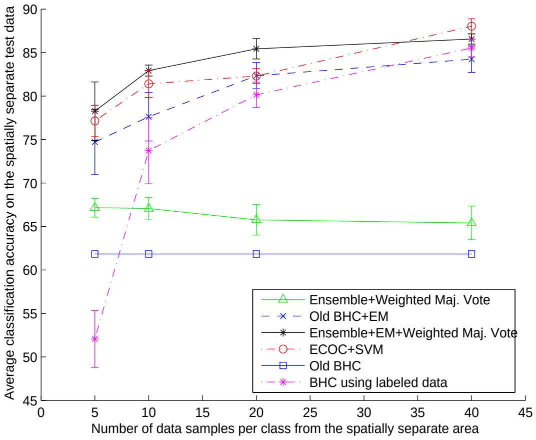

Fig. 1. Average semi-supervised classification accuracies for the KSC dataset.

The Botswana dataset benefits a lot more from the information in area 1 than the KSC. While the area 1 and area 2 data of the Botswana dataset were obtained from the same flightline [4], the area 2 data for the KSC was obtained from a different subset of the flightline [12]. Therefore, the greater disparity in the spectral signatures of the classes between the two areas in the KSC dataset limits the amount of knowledge that can be transferred from one area to another.

Semi-supervised Case: Fig. 1 and Fig. 2 show the learning curves for the KSC and the Botswana dataset when labeled data is available from area 2. It can be observed from Fig.1 and Fig.2 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 both the KSC and the Botswana datasets, an 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.

Similar to the unsupervised scenario, using the data (labeled and unlabeled) from area 2 with EM to update the statistics of the classifiers improved the classification accuracy of the old BHC for the KSC dataset (Fig. 1). For the Botswana dataset the old BHC with EM showed an improvement only when a sufficient number of labeled samples per class (> 10) were available from the spatially separate area. This may be due to the complexity of the classification problem (14 classes as opposed to 10 in the KSC) or due to a poor choice of the labeled data samples. The high values of standard deviation seem to hint at the latter reason.

By adapting the BHC ensemble components via constrained EM some of them 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. As

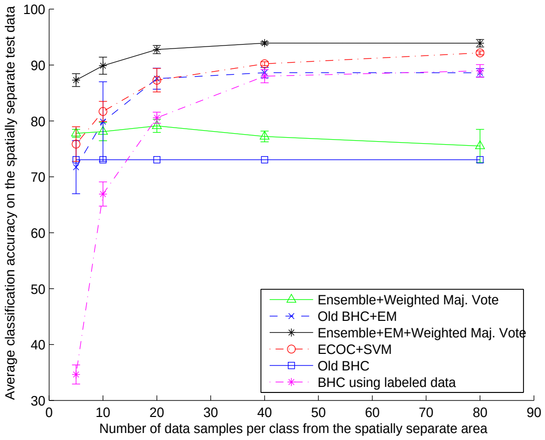

Fig. 2. Average semi-supervised classification accuracies for the Botswana dataset.

more labeled data becomes available from area 2, classifiers trained on that data will perform at least as well if not better than the updated classifiers from area 1. The amount of labeled data from area 2 required for this cross-over is surprisingly a lot especially for the Botswana dataset where knowledge transfer is more effective because of the reason mentioned earlier.

5 Summary and Conclusions

We initially thought that the BHC would be particularly well-suited 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 would be more effective than alternative classifiers, including the maximum likelihood based approach investigated earlier. However, in this application the data characteristics change fairly substantially from area to area, demanding more elaborate 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 (i) using the weighted majority combining approach on an ensemble of trees so that trees more suitable for the new area get higher weights, and (ii) using constrained, semi-supervised EM that can adjust the feature spaces as well as classification boundaries based on both labeled/unlabeled data acquired from the new area. Against this combination, the alternative of building a new classifier from scratch using a powerful and suitable method (ECOC-SVM) was advantageous only when over 30 labeled samples/class were available from the new area for KSC. For Botswana, where the two areas are more similar, the composite knowledge transfer approach was superior even when 80 labeled samples/class were available from the new area. 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.

Acknowledgements: This research was supported by NSF (Grant IIS-0312471).

References

1. A. Blum. On-line algorithms in machine learning. In Fiat and Woeginger, editors, *Online Algorithms: The State of the Art.* LNCS Vol.1442, Springer, 1998.
2. L. Bruzzone and D. F. Prieto. Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE Trans. Geoscience and Remote Sensing*, 39:456–460, 2001.
3. T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.
4. J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Trans. Geoscience and Remote Sensing*, page to appear, 2005.
5. T. Joachims. Making large-scale SVM learning practical. In C. Burges B. Scholkopf and A. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 169–184. MIT Press, Cambridge, USA, 1999.
6. M. G. Kelly, D. J. Hand, and M. N. Adams. The impact of changing populations on classifier performance. In *The Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 67–371, 1999.
7. S. Kumar, J. Ghosh, and M. M. Crawford. Best-bases feature extraction algorithms for classification of hyperspectral data. *IEEE Trans. Geoscience and Remote Sensing*, 39(7):1368–79, 2001.
8. S. Kumar, J. Ghosh, and M. M. Crawford. Hierarchical fusion of multiple classifiers for hyperspectral data analysis. *Pattern Analysis and Applications, spl. Issue on Fusion of Multiple Classifiers*, 5(2):210–220, 2002.
9. L. I. Kuncheva. Classifier ensembles for changing environments. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, pages 1–15. LNCS Vol. 3077, Springer, 2004.
10. L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles. *Machine Learning*, 51:181–207, 2003.
11. N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.
12. J. T. Morgan. *Adaptive Hierarchical Classifier with Limited Training Data.* PhD thesis, Dept. of Mech. Eng., Univ. of Texas at Austin, 2002.
13. 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". *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 18:777–800, 2004.
14. K. Nigam, A. K. Mccallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.
15. S. Rajan and J. Ghosh. An empirical comparison of hierarchical vs. two-level approaches to multiclass problems. In F. Roli, J. Kittler, and T. Windeatt, editors, *Multiple Classifier Systems*, pages 283–292. LNCS Vol. 3077, Springer, 2004.
16. 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. Geoscience and Remote Sensing*, 32:1087–1095, 1994.