MULTI-TEMPORAL REMOTE SENSING IMAGE CLASSIFICATION - A MULTI-VIEW APPROACH

VARUN CHANDOLA* AND RANGA RAJU VATSAVAI*

Abstract. Multispectral remote sensing images have been widely used for automated land use and land cover classification tasks. Often thematic classification is done using single date image, however in many instances a single date image is not informative enough to distinguish between different land cover types. In this paper we show how one can use multiple images, collected at different times of year (for example, during crop growing season), to learn a better classifier. We propose two approaches, an ensemble of classifiers approach and a co-training based approach, and show how both of these methods outperform a straightforward *stacked vector* approach often used in multi-temporal image classification. Additionally, the co-training based method addresses the challenge of limited labeled training data in supervised classification, as this classification scheme utilizes a large number of unlabeled samples (which comes for free) in conjunction with a small set of labeled training data.

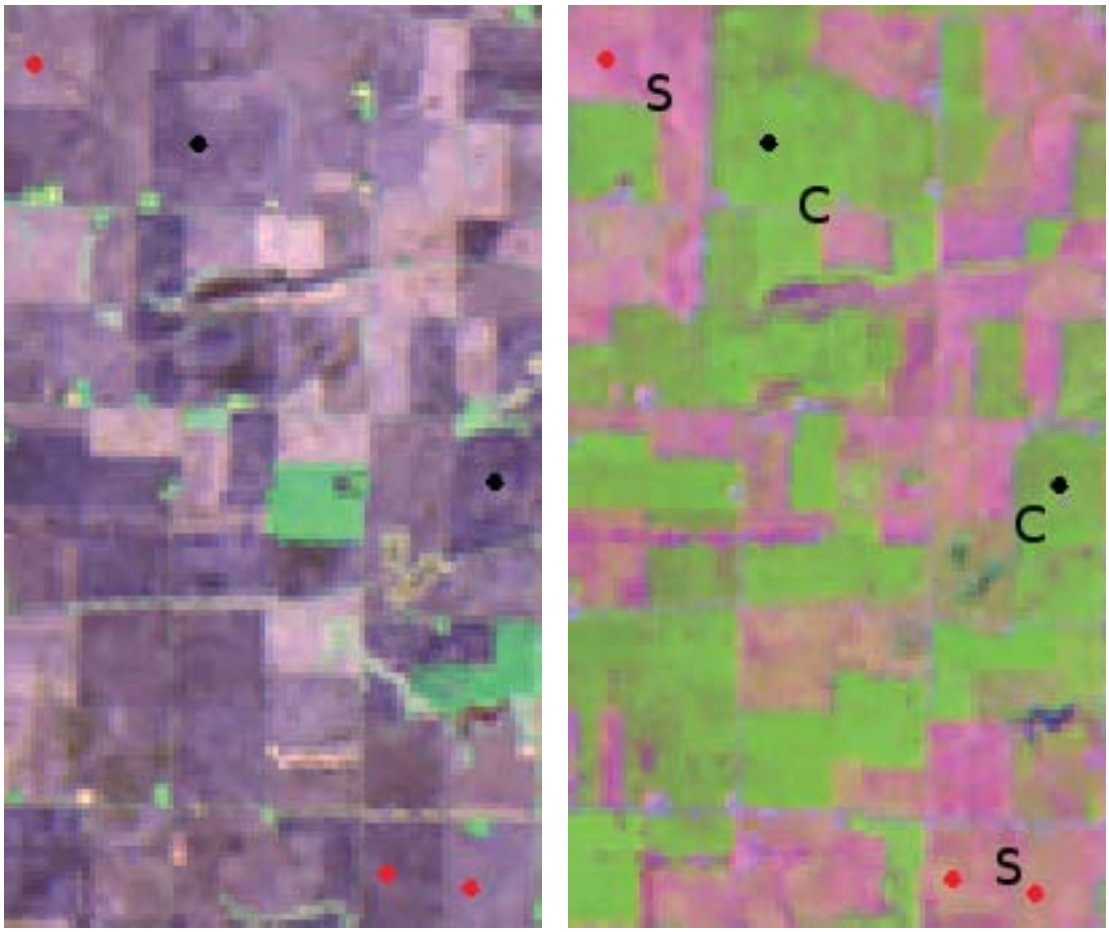
1. Introduction

Multispectral images collected by remote sensing instruments present an immense opportunity for understanding the dynamic characteristics of the earth surface. In the last couple of decades land use and land cover (LULC) identification with remotely sensed images has become of great interest to researchers from various disciplines including earth scientists and data miners, and it has been applied to a variety of applications such as urban planning, natural resource management, water resource monitoring, environmental and agricultural analyses. Remotely sensed multispectral imaging is one of the most widely used technologies for LULC mapping and monitoring, and it provides synoptic and timely information over large geographical areas.

Thematic classification is the most widely used technique for extracting useful and interesting patterns from remote sensing imagery. Several classification algorithms have been proposed in the literature for analysis of remote sensing imagery. These algorithms can be broadly grouped into two categories, supervised and unsupervised, based on the learning scheme used. Among supervised classification methods, the maximum likelihood classifier (MLC) is the most extensively studied and utilized for classifications of multi-spectral images. Other broad classification schemes are neural networks, decision trees, and support vector machines. Among unsupervised methods, the K-Means, C-Means (also known as Migrating Means or ISODATA) and Fuzzy C-Means techniques are popular in remote sensing. Most of these methods work well if the land cover classes are spectrally separable. In reality, the classes under investigation are often spectrally overlapping as the reflectance from these classes is dependent on several extraneous factors like terrain, soil type, moisture content, acquisition time, atmospheric conditions, etc. Though such factors can be incorporated into classification via ancillary data, spectral overlapping due to temporal nature of classes can be separated by the utilization of multi-temporal images. As an illustration we show two images, one taken in May and the other acquired in July. Figure 1 shows how two thematic classes, Soybean (three red plots) and corn (two black plots), which are highly overlapping (meaning, the class spectral reflectances are highly similar) in May (all 5 plots are almost same indigo color) are spectrally dissimilar in July (corn is greenish and soybean is purplish – thus easy to separate). Though Corn and Soybean can be easily separated in June, there may be other classes which are not easily separable in July but

*Oak Ridge National Laboratory, chandolav@ornl.gov, vatsavairr@ornl.gov.

may be separable in May or some other date. This is the basic motivation for multi-temporal image classification, where one seeks to accurately classify thematic classes which are highly overlapping in any single date image.



(a) AWiFS May 3, 2008, FCC (RGB Bands 4, 3, 2), Thematic Classes (C-Corn, S-Soybean)

(b) AWiFS July 14, 2008, FCC (RGB Bands 4, 3, 2), Thematic Classes (C-Corn, S-Soybean)

FIGURE 1. False color composite (FCC) images of same location at two different dates

The rest of this paper is organized as follows: Section 2 describes the related work. Section 3 presents background information on learning with multiple views and section 4 provides basic notions used. In section 5, we describe the maximum likelihood classification framework that provides backbone for Bayesian model averaging (described in Section 6) and co-training (Section 7). Datasets used in this study are described in Section 8 followed by the results and comparative analysis of various classification schemes in Section 9. Finally, conclusions and future directions are provided in Section 10.

2. Related Work

Several studies have used machine learning tools such as decision trees [9, 10] and support vector machines (SVM) [22, 4, 16, 2, 18] to build a multi-class classifier for crop classification using multispectral remote sensing data as well as explored methods to extract features to enhance the classification performance [14, 18]. Such methods typically deal with a single multispectral image.

However, these methods can be readily applied to multi-temporal images by combining all bands (features) – an approach known as *stacked vector*. Though, stacked vector approach do not require any modification to existing approaches, increasing number of features require addition ground truth data which is often costly to obtain. Typically one needs 10-30 times d (d - number of dimensions) samples for accurate fitting of the learning model [15]. Therefore, multi-temporal image classification requires careful design and should not increase the need for additional training data.

In contrast, several papers have used the time series of spectral observations collected across a temporal span, as a data instance for every location[6, 13, 8, 5]. Typically, such approaches do not use the entire spectrum but use a single composite observation, such as *Normalized Difference Vegetation Index* (NDVI), to construct a univariate time series at each spatial location.

The multi-temporal image classification approach proposed in [12], is based on"decision fusion", where a classification model was built separately on each image, and the decisions (predictions) combined using two different fusion criteria. Though our proposed approaches are conceptually similar to the above method, the co-training method substantially differs in two respects: first, it does not fuse the independent classification decisions in the end as with the other methods; second, it incorporates unlabeled training samples, thus offers a more cost-effective solution for multi-temporal image classification.

3. Learning with Multiple Views - Background

In this paper, we treat multi-temporal images as multiple views of same phenomena under study. There are four broad approaches to learn a classifier from data described using multiple views. The first approach is to simply train a classifier on a single view which gives best performance. The choice of the best view can be either made using domain knowledge or through empirical evaluation.

For the second approach, also known as the *stacked vector* approach, feature vectors from all views are concatenated together to get a single composite view of the data. The stacked vector approach results in a increase in the dimensionality of the data.

The third approach is to learn individual classifiers using each view of the data and then combine the predictions of the individual classifiers. Such classification methods are also broadly referred to as *multiple classifier systems* [1, 21, 17, 20, 7][1]. *Bayesian Model Averaging* (BMA) [11, 7] is a probabilistic method for combining the output of multiple classifiers. We describe this method in more detail in Section 6.

The fourth approach has been developed in the context of semi-supervised learning, i.e., using a small set of labeled data and a larger set of unlabeled data. One of the earliest work in this direction was proposed by Blum and Mitchell [3], known as *co-training*. The authors assume that each data instance can be described using two disjoint sets of features, such that each feature set is sufficient for learning, given enough labeled data. In the co-training framework, the key idea is to learn a classifier on each view of the data independently, and then use the predictions of each classifier on unlabeled data instances to augment the training data set for the other classifier. By learning in an iterative fashion, the authors argue that the overall classification performance can be improved.

We describe a generalized co-training based algorithm for multi-temporal (multi-view) classification in Section 7.

4. Notation

We first describe the notations used in this paper. Labeled training examples are denoted as $\{(\mathbf{x_i}, y_i)\}_{i=1}^{l}$, such that each example $\mathbf{x}$ is described using v views, i.e., $\mathbf{x} \equiv \langle \mathbf{x^{(1)}}, \mathbf{x^{(2)}}, \ldots, \mathbf{x^{(v)}} \rangle$ and $\mathbf{x^k} \in \Re^d$, for $k = 1 \ldots v$. In this paper, we are concerned with a multi-class classification problem, where $y \in \{c_1, c_2, \ldots, c_k\}$. Additionally, there exist unlabeled training examples, denoted as $\{\dot{\mathbf{x}}_\mathbf{i}\}_{i=1}^{u}$. The labeled training set is also denoted as X and the unlabeled training set is denoted as U.

[1]Note that these are different from *ensemble classification* methods such as *bagging* and *boosting* which learn multiple classifiers using a single view of the data.

Note that in the above notation scheme all views are assumed to be described using d continuous valued features. In general, however, the different views can be defined using different number of features. Moreover, the features are not constrained to be in $\Re$ and can have arbitrary type (categorical, binary, ordinal), as long as the base classifier that uses those features can handle such types. For simplicity, we follow the above stated notation.

5. Maximum Likelihood Classification

All classification approaches investigated in this paper, i.e., single view, stacked vector, Bayesian model averaging, and co-training, require a base classifier. *Maximum Likelihood Classifier* (MLC) is the most widely used method for land cover classification based on multi-spectral remote sensing imagery because of its simplicity and efficiency[19]. Therefore we employed MLC as a base classifier in this research.

A typical maximum likelihood classifier models the class-conditional distribution, $p(\mathbf{x}|y)$ as a multivariate Gaussian distribution:

$$p(\mathbf{x}|y = c_i) \sim N(\mu_i, \Sigma_i) \tag{1}$$

The parameters for the multivariate Gaussian for each class are obtained using maximum likelihood estimation using the labeled training examples. To assign a class label to a test example, $\mathbf{x}^*$, the posterior probability for each class, given the test example, is computed as:

$$P(y^* = c_i|\mathbf{x}^*, \{(\mathbf{x_i}, y_i)\}_{i=1}^l) \propto p(\mathbf{x}^*|y^* = c_i)P(c_i) \tag{2}$$

where $p(\mathbf{x}|y = c_i)$ is computed using (1) and $P(c_i)$ denotes the prior probability for each class. The class with maximum posterior probability is chosen as the predicted class for the test instance, $\mathbf{x}^*$.

The above described MLC algorithm can be directly used for the single view as well as the stacked vector approach to handle the multiple views.

6. Bayesian Model Averaging

The Bayesian model averaging approach [11, 7] combines the output of multiple classifiers to obtain a single decision for an unseen test instance. In the context of this paper, the multiple classifiers are learnt using different views of the data and are represented as $\{\hbar_1, \hbar_2, \ldots, \hbar_v\}$.

According to the BMA approach, the posterior probability for a class c_i is computed as:

$$P(c_i|\mathbf{x}^*, \{(\mathbf{x_i}, y_i)\}_{i=1}^l) = \sum_{j=1}^{v} P(c_i|\mathbf{x}^*, \{(\mathbf{x_i}, y_i)\}_{i=1}^l, \hbar_j) P(\hbar_j|\{(\mathbf{x_i}, y_i)\}_{i=1}^l) \tag{3}$$

where $P(c_i|\mathbf{x}^*, \{(\mathbf{x_i}, y_i)\}_{i=1}^l, \hbar_j)$ is the posterior density obtained for class c_i using the j^{th} view (See (2)). The second term in the right hand side of (3) is the *model posterior* for the j^{th} model, and is computed as:

$$P(\hbar_j|\{(\mathbf{x_i}, y_i)\}_{i=1}^l) \propto P(\hbar_j) \prod_{i=1}^{l} P(\mathbf{x_i}, y_i|\hbar_j) \tag{4}$$

$P(\hbar_j)$ is the *model prior*. Each term in the product in (4) is the joint probability for the training example, $\mathbf{x_i}$, and the true class, y_i, and can be expressed as: $P(\mathbf{x_i}, y_i|\hbar_j) \propto P(y_i|\mathbf{x_i}, \hbar_j)$ which is the posterior probability of class y_i assigned by the classifier $\hbar_j$ (See (2)). Finally, the class assigned to the test instance $\mathbf{x}^*$ is the one for which the posterior in (3) is maximum. Thus the BMA approach assigns more weight to the classifier which assigns high posterior probabilities to the true class for the training examples.

7. Co-training

In this section we present a co-training based algorithm based on the original algorithm proposed by Blum and Mitchell [3]. While originally, co-training was proposed for two views of the data, we propose a generalized version in which data can be defined using more than two views. Algorithm 1 lists the steps for the training part of the co-training algorithm. The output of this algorithm is a set of v classifiers, one for each view.

Input: $(X = \{(\mathbf{x_i}, y_i)\}_{i=1}^{l}$,$U = \{\dot{\mathbf{x}}_\mathbf{i}\}_{i=1}^{u})$,$\delta$
Output: $\{\hbar_j\}_{j=1}^{v}$
Sample m instances without replacement from U into a set $U' = \{\ddot{\mathbf{x}}_\mathbf{i}\}_{i=1}^{m}$
while U is not empty **do**
  foreach $j = 1 : v$ **do**
    Learn MLC $\hbar_j$ using $\{(\mathbf{x_i^{(j)}}, y_i)\}_{i=1}^{l}$
    Assign class label $\ddot{y}_i$ to each $\ddot{\mathbf{x}}_\mathbf{i} \in U'$ using $\hbar_j$
    foreach $i = 1 : m$ **do**
      if $P(\ddot{y}_i|\ddot{\mathbf{x}}_\mathbf{i}, \hbar_j) \geq \delta$ **then**
        Add $(\ddot{\mathbf{x}}_\mathbf{i}, \ddot{y}_i)$ to X
        Sample one instance without replacement from U into U'
      end
    end
  end
end

Algorithm 1: Co-training

The co-training algorithm starts with the labeled training set X and unlabeled training set U. A smaller unlabeled training set, U' is sampled from U. A MLC classifier is learnt using first view of training set X. The classifier then assigns labels to the unlabeled instances in U. The predictions for which the prediction probability is greater than a certain threshold, δ, are added to the labeled training set. In the next step, a classifier is learnt using the second view of the augmented training set. This process is repeated until all unlabeled instances in U are labeled and added to X. The algorithm finally returns the v classifiers trained on individual views of the final training data set X. The threshold δ is used to include only those unlabeled instances to the training data set which are predicted with high probability.

The order in which the views are used in the co-training algorithm is arbitrary. In the above algorithm we use the natural ordering of the views, though experimentally we have observed that the choice of ordering does not have a significant impact on the performance.

For testing, the algorithm follows the same procedure as that of the BMA classifier (See Section 6).

8. Data

This research was carried out in the north-west portion of the Iowa state, U.S.A. The predominant thematic classes in this study areas are corn and soybean. Table 1 shows other thematic classes and the number of labeled samples (plots) collected over different portions of the image. Each training plot size is 3 x 3 window (that is, 9 pixels). The ground truth for training, testing and thematic classes were all based on the crop data layer data produced by the United Stated Department of Agriculture (USDA). The remote sensing images used in this study were acquired on four different dates in 2008: May 03, July 14, August 31, and September 24, by the IRS-P6 satellite using the Advanced Wide Field Sensor (AWiFS) camera. There are four spectral bands in each image with a spatial resolution of 56 meters. Sample image covering 370 x 370 km along with spatial location is shown in Figure 2. For this study we used 3 bands (red, near-infrared, and short-wave infrared)

from each images. Black dots are the sample locations where ground truth (training and testing) data was collected.

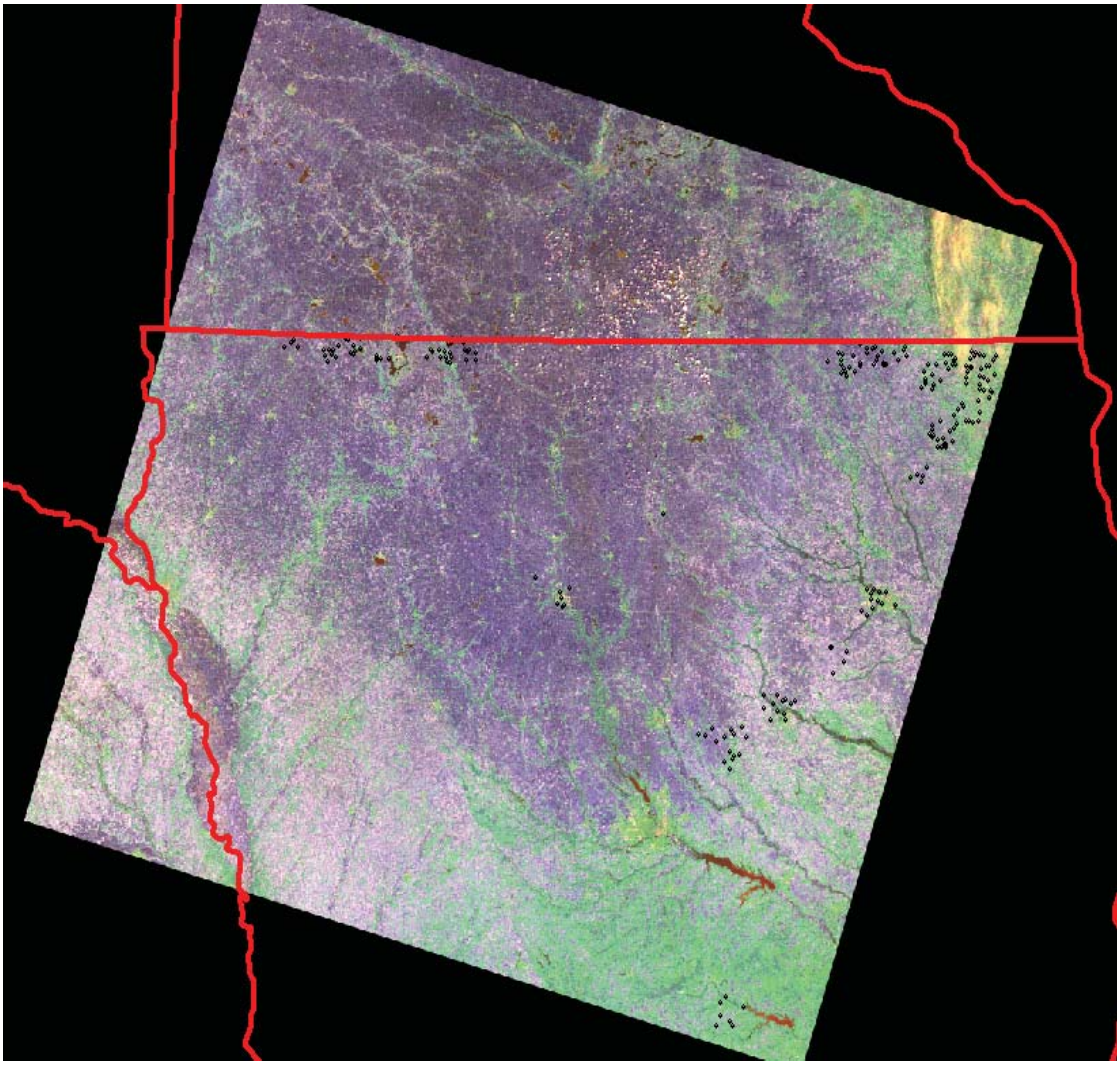


FIGURE 2. False color composite (FCC; Bands 4,3,2) Image Acquired on May 08, 2008 by IRS-P6, AWiFS, over Iowa

9. Results and Analysis

In this section we compare the performance of various classification methods discussed in this paper on MODIS data described in Section 8. MLC is used as the base classifier for all approches. A uniform prior is assumed over all classes in 2. For BMA, a uniform prior is assumed over all classifiers $(\hbar_i, \hbar_2, \ldots, \hbar_v)$ in (4). For co-training, the δ threshold was set to 0.90. We experimentally observed that the performance of the co-training based classifier is not sensitive to δ in the range of $[0.8, 0.95]$. For each classifier we report the following:

(1) Confusion matrix.
(2) Per-class recall, precision, and F-measure.
(3) Misclassification error.

9.1. **Comparing Bayesian Averaging and Stacked Vector Approach.** We first compare the performance of the two supervised methods to handle multiple views of data, i.e., Bayesian averaging and stacked vector approach. For comparative purpose, we also report the performance of a ML classifier using an individual view (image) only. For each of these experiments we trained on labeled data set corresponding to 945 locations and tested on a validation data set corresponding to 963 locations. For each location there are four views, corresponding to four images collected in four different months (May, July, August, September) and each view consists of three spectral bands. The details of the training and validation data sets are summarized in Table 1.

Class ID	Class	Training	Validation
1	Corn	261	261
5	Soybean	225	225
36	Alfa alfa	27	27
62	Grass	189	180
111	Water	18	18
121	Developed	90	99
141	Deciduous Forest	117	117
190	Wetlands Forest	18	36
	Total:	945	963

TABLE 1. Details of Training and Validation Data Set

The confusion matrices obtained from data corresponding to individual views are shown in Tables 2–5, respectively. In all the confusion matrix tables, we also report the per-class recalls in the last column, and the per-class precisions and per-class F-measures in the last two rows of the table, respectively. The last value in the precision row is the fraction of instances that are correctly classified. The last value in the F-measure row is the average F-measure across all classes.

		Predicted								
	Class	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd	Rec_i
Actual	corn	191	45	0	12	0	2	11	0	0.73
	soy	126	96	0	1	0	0	2	0	0.43
	alfa	0	0	18	9	0	0	0	0	0.67
	grass	8	0	16	144	0	7	5	0	0.80
	water	11	0	0	0	4	0	0	3	0.22
	dvlpd	2	9	2	10	0	74	2	0	0.75
	forest	0	0	0	1	0	9	107	0	0.91
	wetlnd	1	0	0	0	0	0	2	33	0.92
	$Prec_i$	0.56	0.64	0.50	0.81	1.00	0.80	0.83	0.92	**0.69**
	F_i	0.64	0.51	0.57	0.81	0.36	0.77	0.87	0.92	**0.68**

TABLE 2. Confusion matrix for MLC on May image only.

In order to understand the overlapping nature of classes in various images and its impact on classification accuracy, we computed pairwise transformed divergence. Transformed divergence is a signature separability measure often used by remote sensing analysts to gain understanding into the class separability in feature space. The formula for transformed divergence T_{ij} between classes i and j is:

$$T_{ij} = 2000(1 - \exp(-\frac{D_{ij}}{8})) \tag{5}$$

where D_{ij} is the divergence between classes i and j, and can be computed as:

$$D_{ij} = \frac{1}{2}tr((\Sigma_i - \Sigma_j)(\Sigma_i^{-1} - \Sigma_j^{-1})) + \frac{1}{2}tr(((\Sigma_i^{-1} - \Sigma_j^{-1}))(\mu_i - \mu_j)(\mu_i - \mu_j)^{\mathsf{T}}) \tag{6}$$

		Predicted								
	Class	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd	Rec_i
Actual	corn	208	11	0	22	0	0	20	0	0.80
	soy	4	202	0	19	0	0	0	0	0.90
	alfa	9	18	0	0	0	0	0	0	0.00
	grass	48	36	0	90	0	6	0	0	0.50
	water	0	0	0	0	18	0	0	0	1.00
	dvlpd	0	4	0	3	0	89	0	3	0.90
	forest	9	0	0	0	0	0	98	10	0.84
	wetlnd	0	0	0	0	0	0	24	12	0.33
	$Prec_i$	0.75	0.75	–	0.67	1.00	0.94	0.69	0.48	**0.74**
	F_i	0.77	0.81	0.00	0.57	1.00	0.92	0.76	0.39	**0.65**

TABLE 3. Confusion matrix for MLC on July image only.

		Predicted								
	Class	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd	Rec_i
Actual	corn	232	3	0	0	0	0	17	9	0.89
	soy	12	186	9	18	0	0	0	0	0.83
	alfa	7	9	9	0	0	0	2	0	0.33
	grass	5	13	21	119	0	11	2	9	0.66
	water	0	0	0	0	18	0	0	0	1.00
	dvlpd	0	0	0	2	0	96	0	1	0.97
	forest	7	0	0	0	0	2	94	14	0.80
	wetlnd	0	0	0	0	0	0	36	0	0.00
	$Prec_i$	0.88	0.88	0.23	0.86	1.00	0.88	0.62	0.00	**0.78**
	F_i	0.89	0.85	0.27	0.75	1.00	0.92	0.70	0.00	**0.67**

TABLE 4. Confusion matrix for MLC on August image only.

		Predicted								
	Class	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd	Rec_i
Actual	corn	171	27	3	19	0	4	29	8	0.66
	soy	12	180	2	11	0	20	0	0	0.80
	alfa	0	0	9	18	0	0	0	0	0.33
	grass	27	22	13	109	0	0	1	8	0.61
	water	0	0	0	0	12	6	0	0	0.67
	dvlpd	9	25	0	0	0	60	5	0	0.61
	forest	8	0	0	0	0	18	66	25	0.56
	wetlnd	0	8	0	0	0	10	18	0	0.00
	$Prec_i$	0.75	0.69	0.33	0.69	1.00	0.51	0.55	0.00	**0.63**
	F_i	0.70	0.74	0.33	0.65	0.80	0.55	0.56	0.00	**0.54**

TABLE 5. Confusion matrix for MLC on September image only.

A transformed divergence value of less than 1500 between two classes indicates that those two classes can't be separated, in other words, there will be lot of misclassification between those two classes. In conjunction with transformed divergence, results of the ML classifier trained on individual views provide several interesting insights:

(1) In May (crop planting season), the corn and soybean crops are not clearly distinguishable, but are clearly separable in later months. Transformed divergence between corn and soy in May is 957.98 (Table 6) which indicates that these two classes are highly overlapping. MLC shows that 45 samples from *corn* are misclassified as *soy* and 126 samples of *soy* are misclassified as *corn*. On the other-hand, a transformed divergence value of 1610.59 (Table 7) indicates that these classes are fairly separable, which is also reflected in MLC performance in July image.

	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd
corn	0.00	957.98	2000.00	1999.98	2000	1999.45	1859.75	2000
soy	957.98	0.00	2000.00	2000.00	2000	2000.00	1999.11	2000
alfa	2000.00	2000.00	0.00	2000.00	2000	1998.70	1999.89	2000
grass	1999.98	2000.00	2000.00	0.00	2000	1790.64	1973.95	2000
water	2000.00	2000.00	2000.00	2000.00	0.00	2000.00	2000.00	2000
dvlpd	1999.45	2000.00	1998.70	1790.64	2000	0.00	1817.02	2000
forest	1859.75	1999.11	1999.89	1973.95	2000	1817.02	0.00	2000
wetlnd	2000.00	2000.00	2000.00	2000.00	2000	2000.00	2000.00	0.00

TABLE 6. Transformed Divergence Between Classes from May Image

	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd
corn	0.00	1610.59	2000	927.95	2000	2000.00	1993.94	1999.65
soy	1610.59	0.00	2000	1252.87	2000	1997.30	2000.00	2000.00
alfa	2000.00	2000.00	0.00	2000.00	2000	2000.00	2000.00	2000.00
grass	927.95	1252.87	2000	0.00	2000	1992.04	1999.50	1999.76
water	2000.00	2000.00	2000	2000.00	0.00	2000.00	2000.00	2000.00
dvlpd	2000.00	1997.30	2000	1992.04	2000	0.00	2000.00	1999.31
forest	1993.94	2000.00	2000	1999.50	2000	2000.00	0.00	1734.34
wetlnd	1999.65	2000.00	2000	1999.76	2000	1999.31	1734.34	0.00

TABLE 7. Transformed Divergence Between Classes from July Image

(2) Likewise one can see in Table 7 that grass in July image is confusing with corn and soy classes, however they are fairly separable in May image.
(3) Wetlands are better identified when using May data but are completely missed by classifiers that use August and September data.
(4) The classifier that uses May data performs poorly in identifying water, but the classifiers using data from later months perform significantly better for water.

		Predicted								
	Class	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd	Rec_i
Actual	corn	252	0	0	2	0	0	7	0	0.97
	soy	0	224	0	1	0	0	0	0	1.00
	alfa	0	0	0	27	0	0	0	0	0.00
	grass	9	0	0	170	0	1	0	0	0.94
	water	0	0	0	0	0	18	0	0	0.00
	dvlpd	0	0	0	0	0	99	0	0	1.00
	forest	4	0	0	3	0	0	110	0	0.94
	wetlnd	14	0	0	2	0	2	18	0	0.00
	$Prec_i$	0.90	1.00	–	0.83	–	0.82	0.81	–	**0.89**
	F_i	0.93	1.00	0.00	0.88	0.00	0.90	0.87	0.00	**0.57**

TABLE 8. Confusion matrix for the stacked vector method.

The confusion matrices for the classifiers trained using the stacked vector and Bayesian averaging classifier are shown in Tables 8 and 9, respectively. On average, both of these methods perform better than the classifiers trained using individual views. This is expected, since data collected from different months have distinguishing abilities for different types of land cover. The stacked vector method classifies 89% of instances correctly, but the F-measure reveals that it completely misses the smaller classes, like alfa-alfa, water, and wetlands. The reason for this is that the dimensionality of

		Predicted								
	Class	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd	Rec_i
Actual	corn	232	3	0	0	0	0	17	9	0.89
	soy	12	186	9	18	0	0	0	0	0.83
	alfa	7	9	9	0	0	0	2	0	0.33
	grass	5	13	21	119	0	11	2	9	0.66
	water	0	0	0	0	18	0	0	0	1.00
	dvlpd	0	0	0	2	0	96	0	1	0.97
	forest	7	0	0	0	0	2	94	14	0.80
	wetlnd	0	0	0	0	0	0	36	0	0.00
	$Prec_i$	0.88	0.88	0.23	0.86	1.00	0.88	0.62	0.00	**0.78**
	F_i	0.89	0.85	0.27	0.75	1.00	0.92	0.70	0.00	**0.77**

TABLE 9. Confusion matrix for the Bayesian averaging method.

the input is large (12) and hence the parameter estimation for the smaller classes is inaccurate (also known as the Hughes effect). Since the Bayesian averaging method learns classifiers for individual views, it does not get affected by the high-dimensionality issue and hence performs better on small classes. Since the Bayesian averaging method combines the classifiers trained on individual views, it is able to perform better than the individual classifiers, though it cannot correctly identify any of the instances belonging to the wetlands class.

9.2. **Comparing Co-training with Supervised Multi-view Learning Approaches.** In this section we present results using the co-training method. Since co-training is a semi-supervised learning approach we use a small fraction of the available labeled training data for training. The remaining training instances are used as the unlabeled data used by the co-training algorithm. The labeled instances are picked randomly. We experimented with 10 different random samples and report the average results of the 10 resulting confusion matrices. Table 10 shows the confusion matrix obtained for the co-training approach using a labeled data set of size 120. Table 11 shows the confusion matrix when the size of the labeled data set was 400.

		Predicted								
	Class	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd	Rec_i
Actual	corn	245	6	0	1	0	0	9	0	0.94
	soy	14	209	1	1	0	0	0	0	0.93
	alfa	0	0	18	9	0	0	0	0	0.67
	grass	10	0	17	136	0	12	0	5	0.76
	water	0	0	0	0	18	0	0	0	1.00
	dvlpd	0	3	0	1	0	94	1	0	0.95
	forest	6	0	0	0	0	0	108	3	0.92
	wetlnd	0	0	0	0	0	0	23	13	0.36
	$Prec_i$	0.89	0.96	0.50	0.92	1.00	0.89	0.77	0.62	**0.87**
	F_i	0.91	0.94	0.57	0.83	1.00	0.92	0.84	0.46	**0.81**

TABLE 10. Confusion matrix for co-training using 120 labeled training instances.

We immediately notice from Table 10 that the co-training based method uses only 120 labeled training instances and still significantly outperforms the stacked vector and Bayesian averaging based classifiers which use 945 labeled training instances. Increasing the number of training instances for co-training to 400 only marginally improves the performance. Moreover, the co-training classifier performs well on all classes, even those for which other classifiers performed poorly, like alfa-alfa, water, and wetland. The key strength of co-training is that it iteratively adds high quality unlabeled instances to the training set and hence builds classifiers (for each view) using a relatively higher quality training data compared to the entire data set used by the other methods.

For comparison we also report the performance of the stacked vector and the Bayesian model averaging methods using the same labeled training data set (of size 120) as used by the co-training

		Predicted								
	Class	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd	Rec_i
Actual	corn	244	8	0	0	0	0	9	0	0.93
	soy	15	209	0	1	0	0	0	0	0.93
	alfa	1	0	10	16	0	0	0	0	0.37
	grass	10	0	7	146	0	10	1	7	0.81
	water	0	0	0	0	18	0	0	0	1.00
	dvlpd	0	2	0	1	0	96	0	0	0.97
	forest	3	0	0	0	0	0	113	1	0.97
	wetlnd	0	0	0	0	0	0	26	11	0.30
	$Prec_i$	0.89	0.95	0.59	0.89	1.00	0.91	0.76	0.58	**0.88**
	F_i	0.91	0.94	0.45	0.85	1.00	0.94	0.85	0.39	**0.79**

TABLE 11. Confusion matrix for co-training using 400 labeled training instances.

algorithm. This was done to ensure that the subset of 120 instances, by itself is not enough to learn a good classifier. Tables 12 and 9.2 show that the performance of these classifiers significantly deteriorates compared to when the larger training data is used (Tables 8 and 9). This indicates that the iterative augmentation of training data by co-training is indeed a better way to incorporate multiple views of the data as well as unlabeled training instances.

		Predicted								
	Class	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd	Rec_i
Actual	corn	231	0	0	28	0	0	2	0	0.89
	soy	28	162	0	35	0	0	0	0	0.72
	alfa	0	0	0	27	0	0	0	0	0.00
	grass	0	0	0	180	0	0	0	0	1.00
	water	4	0	0	5	0	9	0	0	0.00
	dvlpd	17	0	0	53	0	29	0	0	0.29
	forest	6	0	0	70	0	0	41	0	0.35
	wetlnd	6	0	0	19	0	5	6	0	0.00
	$Prec_i$	0.79	1.00	–	0.43	–	0.67	0.84	–	**0.67**
	F_i	0.84	0.84	0.00	0.60	0.00	0.41	0.49	0.00	**0.40**

TABLE 12. Confusion matrix for stacked vector method using 120 labeled training instances.

		Predicted								
	Class	corn	soy	alfa	grass	water	dvlpd	forest	wetlnd	Rec_i
Actual	corn	212	12	0	20	0	0	15	3	0.81
	soy	11	194	4	15	0	0	0	0	0.87
	alfa	8	15	3	1	0	0	0	0	0.11
	grass	29	29	6	105	0	8	0	2	0.59
	water	0	0	0	0	18	0	0	0	1.00
	dvlpd	1	6	0	4	0	83	0	5	0.84
	forest	12	0	0	1	0	1	83	21	0.70
	wetlnd	0	0	0	0	0	0	18	18	0.50
	$Prec_i$	0.78	0.76	0.23	0.72	1.00	0.90	0.72	0.37	**0.74**
	F_i	0.79	0.81	0.15	0.65	1.00	0.87	0.71	0.42	**0.67**

TABLE 13. Confusion matrix for Bayesian averaging method using 120 labeled training instances.

10. CONCLUSIONS

In this paper we proposed two approaches for classifying multi-temporal images. In the first approach, we used fusion of predictions from ensemble of classifiers using Bayesian model averaging.

In the second approach we generalized co-training method for multiple views. We compared the performance of these two classification schemes with regular MLC and straightforward *stacked vector* approach that are often used in multi-temporal image classification. All four methods were evaluated on multi-temporal images from four different dates spanning crop growing season in 2008. Evaluation on independent test dataset shows the better overall performance of co-training based method over all three other methods. The key strength of co-training is that it iteratively adds high quality unlabeled instances to the training set and hence builds classifiers (for each view) using a relatively higher quality training data compared to the entire data set used by the other methods. As co-training requires less number of labeled samples as compared to the other methods, this methods can be widely used in multi-temporal image classification over large geographic regions.

11. Acknowledgments

Prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UT-Battelle, LLC for the U. S. Department of Energy under contract no. DEAC05-00OR22725. This research is funded through the LDRD program at ORNL.

References

[1] J. A. Benediktsson, J. Kittler, and F. Roli, editors. *Proceedings of 5th International Workshop on Multiple Classifier Systems*, volume 3077 of *Lecture Notes in Computer Science*. Springer, 2009.

[2] J. A. Benediktsson, Y. Tarabalka, B. Waske, M. Fauvel, and J. R. Sveinsson. Ensemble methods for classication of hyperspectral data. In *IEEE International Geoscience and Remote Sensing Symposium, IGARSS*, 2008.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pages 92–100, New York, NY, USA, 1998. ACM.

[4] G. Camps-Valls, L. Gómez-Chova, J. Calpe-Maravilla, E. Soria-Olivas, J. D. Martón-Guerrero, and J. Moreno. Support vector machines for crop classification using hyperspectral data. *Pattern Recognition and Image Analysis*, 2652:134–141, 2003.

[5] C. Conrad, S. Fritsch, J. Zeidler, G. Rcker, and S. Dech. Per-field irrigated crop classification in arid central asia using spot and aster data. *Remote Sensing*, 2(4):1035–1056, 2010.

[6] R. S. Defries and J. R. G. Townshend. NDVI-derived land cover classifications at a global scale. *International Journal of Remote Sensing*, 15(17):3567–3586, 1994.

[7] P. Domingos. Bayesian averaging of classifiers and the overfitting problem. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 223–230, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[8] P. Doraiswamy, B. Akhmedov, and A. Stern. Crop classification in the u.s. corn belt using MODIS imagery. In *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2007.

[9] M. Friedl and C. Brodley. Decision tree classification of land-cover from remotely-sensed data. *Remote Sensing of Environment*, 61(3):399–409, September 1997.

[10] R. S. D. Fries, M. Hansen, J. R. G. Townshend, and R. Sohlberg. Global land cover classifications at 8 km spatial resolution: the use of training data derived from landsat imagery in decision tree classifiers. *International Journal of Remote Sensing*, 19(16):3141–3168, 1998.

[11] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

[12] B. Jeon, D. A. Landgrebe, and D. A. L. Decision fusion approach for multitemporal classification. *IEEE Transactions on Geoscience and Remote Sensing*, 37:1227–1233, 1999.

[13] J. F. Knight and R. S. Lunetta. Regional scale land cover characterization using MODIS NDVI 250 m multi-temporal imagery: A phenology-based approach. *GIScience and Remote Sensing*, 43(1):1–23, 2006.

[14] S. Mader, M. Vohland, T. Jarmer, and M. Casper. Crop classification with hyperspectral data of the hymap sensor using different feature extraction techniques. In *Proceedings of the 2nd Workshop of the EARSeL Special Interest Group on Land Use and Land Cover*, pages 28–30, 2006.

[15] P. M. Mather. *Computer processing of remotely-sensed images: an introduction*. John Wiley & Sons, Inc., New York, NY, USA, 1988.

[16] A. Mathur and G. M. Foody. Crop classification by support vector machine with intelligently selected training data for an operational application. *International Journal of Remote Sensing*, 29(8):2227–2240, 2008.

[17] O. Okun and H. Priisalu. Multiple views in ensembles of nearest neighbor classifiers. In *Workshop on Learning with Multiple Views, Proceedings of International Conference on Machine Learning*, 2005.
[18] J. Plaza, A. J. Plaza, and C. Barra. Multi-channel morphological profiles for classification of hyperspectral images using support vector machines. *Sensors*, 9(1):196–218, 2009.
[19] J. A. Richards and X. Jia. *Remote Sensing Digital Image Analysis: An Introduction*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.
[20] S. Rüping. Classification with local models. In K. Morik, J.-F. Boulicaut, and A. Siebes, editors, *Proceedings of the Dagstuhl Workshop on Detecting Local Patterns*, 2005.
[21] I. Tsochantaridis and T. Hofmann. Support vector machines for polycategorical classification. In *ECML '02: Proceedings of the 13th European Conference on Machine Learning*, pages 456–467, London, UK, 2002. Springer-Verlag.
[22] J. Zhang, Y. Zhang, and T. Zhou. Classification of hyperspectral data using support vector machine. In *IEEE International Conference on Image Processing*, pages 882–885, 2001.