

Maximizing Land Cover Classification Accuracies Produced by Decision Trees at Continental to Global Scales

Mark A. Friedl, Carla E. Brodley, and Alan H. Strahler, *Member, IEEE*

Abstract—Classification of land cover from remotely sensed data at continental to global scales requires sophisticated algorithms and feature selection techniques to optimize classifier performance. We examine methods to maximize classification accuracies using decision trees to map land cover from multi-temporal AVHRR imagery at continental and global scales. As part of our analysis we test the utility of “boosting,” a new technique developed to increase classification accuracy by forcing the learning (classification) algorithm to concentrate on those training observations that are most difficult to classify. Our results show that boosting consistently reduces misclassification rates by \approx 20–50% depending on the data set in question, and that most of the benefit gained by boosting is achieved after seven boosting iterations. We also assess the utility of including phenological metrics and geographic position as additional features to the classification algorithm. We find that using derived phenological metrics produces little improvement in classification accuracy relative to using an annual time series of NDVI data, but that geographic position provides substantial power for predicting land cover types at continental and global scales. However, in order to avoid generating spurious classification accuracies using geographic position, training data must be distributed evenly in geographic space.

Index Terms—Classification, decision trees, land cover.

I. INTRODUCTION

REMOTE sensing studies of the Earth’s terrestrial ecosystems have witnessed a significant expansion of analysis scale over the past fifteen years [2], [7], [14], [24]. This shift reflects the increased level of interest in global change processes and has prompted new questions and technical issues associated with processing coarse resolution multitemporal data. With the imminent launch of the AM platform of the Earth Observing System (EOS), the need for improved algorithms to process global scale data sets is even more pressing. In this paper, we examine issues associated with supervised classification of coarse spatial resolution data for land cover mapping applications. The motivation for this work is driven by the requirements of global land cover mapping algorithms being developed for use with data from

the Moderate Resolution Imaging Spectroradiometer (MODIS) [27].

The specific objectives of this work are to evaluate two strategies designed to maximize land cover classification accuracies derived from supervised classification algorithms. The first strategy is a new technique known as “boosting” that has recently been developed in the field of machine learning [5]. The second strategy is to supplement time series normalized difference vegetation index (NDVI) measurements with other input features including geographic position and phenological metrics designed to capture dynamics in vegetation.

To assess the utility of these methods, we performed a set of analyses using decision trees to classify two data sets of composited NDVI data. Our results show that boosting is effective for land cover classification problems, but that phenological metrics do not significantly improve classification accuracies relative to classifications based upon a complete twelve month cycle of NDVI measurements. Further we find that while geographic position provides a useful predictor that complements remotely sensed input features, representative training data must be included from each region to be classified in order to avoid spurious results from cross validated estimates of classification accuracy derived from random splits of training and testing data.

II. SUPERVISED CLASSIFICATION AT CONTINENTAL TO GLOBAL SCALES

Virtually all remote sensing studies of land cover and land cover change at continental to global scales have used data from the advanced very high resolution radiometer (AVHRR) on board the NOAA series of meteorological satellites. This instrument provides measurements at sufficiently coarse spatial resolution (1.1 km at nadir) to allow processing and analysis at continental scales. At the same time, relative to higher spectral resolution sensors such as the Landsat Thematic Mapper, the spectral information content provided by the AVHRR is substantially less useful for land cover classification problems.

To compensate for the lower spectral resolution of AVHRR data, the temporal domain has been widely exploited using maximum value NDVI composite data with compositing periods ranging from ten days to one month [10]. Using time series of composited NDVI data, numerous researchers have explored large scale patterns in vegetation, and by extension, land cover type (e.g., [15]). These studies have demonstrated that a wealth of information related to vegetation phenology at

Manuscript received October 7, 1997; revised May 11, 1998. This work was supported in part by NASA Grant NAS5-31369 and IBM Grant 671-1285-2757.

M. A. Friedl and A. H. Strahler are with the Department of Geography and Center for Remote Sensing, Boston University, Boston, MA 02215-1401 USA.

C. E. Brodley is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA.

Publisher Item Identifier S 0196-2892(99)01984-1.

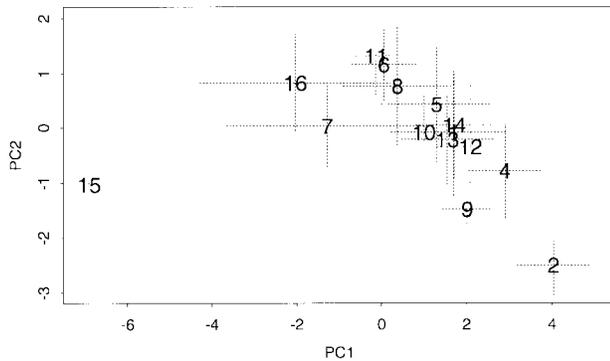


Fig. 1. IGBP class means for the first two principal components estimated from 12 months of 1 km NDVI data over North America ± 1 standard deviation. The numbers on the plot refer to IGBP class values (see Table I).

continental scales can be extracted from time series of AVHRR NDVI measurements [11], [12], [15], [29].

Despite substantial success in extracting information related to vegetation phenology from multitemporal AVHRR data, efforts to map land cover using automated classification algorithms have proven to be more difficult [3]. In particular, the use of supervised classification algorithms in association with multitemporal AVHRR data is fraught with problems related to the separability of classes in spectral-temporal space at continental scales, and the lack of generality in spectral classes derived from training data extracted from a limited number of training sites distributed over continental scales.

To illustrate, Fig. 1 plots class means \pm one standard deviation computed from the first two principal components (representing about 91% of the total variance) of a twelve month time series of 1 km NDVI data from AVHRR over North America. Each number in the figure corresponds to the mean value in each principal component for one of the seventeen classes in the global land cover classification system defined by the International Geosphere-Biosphere Program (IGBP) [14] (see Table I for the class name corresponding to each number). This plot illustrates that substantial overlap exists in the spectral-temporal space of the different IGBP classes (e.g., classes 1, 10, 12–14). Given this overlap, a central issue confronting land cover mapping activities planned under EOS is how to optimize supervised classification algorithms such that these classes can be accurately discriminated and mapped in an efficient and repeatable fashion at continental and global scales.

It is important to note that the strategy being developed to map land cover using EOS data employs a supervised classification model [27]. The choice of a supervised approach is based on the need for automated and repeatable algorithms in order to produce quarterly land cover maps in a timely fashion. While previous studies using AVHRR data in association with supervised classification techniques have proven to be moderately successful in this regard, the improved spectral resolution and radiometric quality of MODIS will provide superior spectral information to complement the temporal information currently being exploited in AVHRR data. At the same time, problems associated with signature extension may substantially complicate the use of supervised classifi-

TABLE I
NDVI 1 km NORTH AMERICA LAND COVER CLASSES
(N = NUMBER OF SAMPLES IN EACH CLASS)

Class Name	N
1 Evergreen needleleaf forest	958
2 Evergreen broadleaf forest	91
3 Deciduous needleleaf forest	0
4 Deciduous broadleaf forest	265
5 Mixed forest	709
6 Closed shrublands	213
7 Open shrublands	539
8 Woody savannas	311
9 Savannas	12
10 Grasslands	425
11 Permanent wetlands	87
12 Cropland	469
13 Urban and built-up	17
14 Cropland/Natural vegetation mosaic	341
15 Snow and ice	597
16 Barren or sparsely vegetated	511
17 Water bodies	0

cation algorithms planned as part of EOS using MODIS data. Within this framework, improved classification algorithms that are robust with respect to noise in training data, improved understanding of the best features available to discriminate among land cover classes, and the development of methods to minimize problems caused by confusion among spectral classes will maximize the quality of land cover maps produced from MODIS data using supervised classification algorithms. In the sections below, we consider these questions using two AVHRR data sets:

- 1 km spatial resolution for North America;
- 1° spatial resolution that includes all land masses on the Earth's surface.

III. DECISION TREES AND BOOSTING—BASIC THEORY

Recent work has demonstrated that decision trees provide an accurate and efficient methodology for land cover classification problems in remote sensing [6], [9], [28]. At global scales, decision trees have recently been used to map land cover using the 8 km AVHRR pathfinder data set with encouraging success [3]. Among the advantages of decision trees that are particularly useful for remote sensing problems are their ability to handle noisy and missing data [22], [25]. Further, they require no assumptions regarding the distribution of input data and also provide an intuitive classification structure.

A. Estimating Decision Trees from Training Data

For this work, we use C5.0, a univariate decision tree algorithm that is the commercial successor of C4.5, a widely used and tested classification algorithm. A complete description of this algorithm is beyond the scope of this paper, and the reader is referred to [22] for complete details. Here we summarize the key components of this algorithm as described in [22], focusing particular attention to those aspects that pertain to estimation of splitting rules and feature selection.

The most important element of a decision tree estimation algorithm is the method used to estimate splits at each internal

node of the tree. To do this, C5.0 uses a metric called the *information gain ratio*, which measures the reduction in entropy in the data produced by a split. Using this metric, the test at each node within a tree is selected using the subdivision of the data that maximizes the reduction in entropy of the descendant nodes. Given a training data set D composed of observations belonging to one of m classes $\{C_1, C_2, \dots, C_m\}$, we desire a test T that partitions D into n mutually exclusive subsets $\{S_1, S_2, \dots, S_n\}$. If we define $f(C_i, D)$ to be equal to the number of cases in D belonging to class C_i , and $|D|$ to be equal to the total number of observations in D , then the amount of information required to identify the class for an observation in D may be quantified as

$$\text{info}(D) = - \sum_{j=1}^m \frac{f(C_j, D)}{|D|} \times \log_2 \frac{f(C_j, D)}{|D|}. \quad (1)$$

Given a test, T , that partitions D into k outcomes $\{D_1, D_2, \dots, D_k\}$, a similar measure may be defined that quantifies the total information content after applying T

$$\text{info}_T(D) = \sum_{i=1}^k \frac{|D_i|}{|D|} \times \text{info}(D_i). \quad (2)$$

Using this approach, we can measure the information gained by splitting D using T as

$$\text{gain}(T) = \text{info}(D) - \text{info}_T(D). \quad (3)$$

The so-called ‘‘gain criteria’’ selects the test for which $\text{gain}(T)$ is maximum. Unfortunately, $\text{gain}(T)$ tends to favor tests with large numbers of splits. To compensate for this effect, $\text{gain}(T)$ is normalized by

$$\text{split info}(T) = - \sum_{i=1}^n \frac{|D_i|}{|D|} \times \log_2 \left(\frac{|D_i|}{|D|} \right) \quad (4)$$

obtaining the splitting metric

$$\text{gain ratio}(T) = \text{gain}(T) / \text{split info}(T). \quad (5)$$

Using this framework, D is recursively split such that the gain ratio is maximized at each node of the tree. This procedure continues until each leaf node contains only observations from a single class or no gain in information is yielded by further splitting.

The result from this procedure is often a very large and complex tree that may be overfit to noise in the training data. If the training data contain errors, then overfitting the tree to the data in this manner can lead to poor performance on unseen cases. To minimize this problem, the original tree must be pruned to reduce classification errors when data outside of the training set are to be classified. To address this problem C5.0 uses error-based pruning. For details, the reader is referred to [18], [21], [22].

B. Boosting

As part of our analysis using decision trees, we test a new technique known as boosting that has recently been developed in the machine learning research community. The goal of boosting is to improve the classification accuracy of a given base or ‘‘weak’’ learning algorithm (i.e., one that provides less than acceptable classification accuracies) [26]. To do this, boosting algorithms estimate multiple classifications in an iterative fashion using the base classification algorithm (in this case C5.0). At each iteration, a weight is assigned to each training observation. Those observations that were misclassified in the previous iteration are assigned a heavier weight in the next iteration, thereby forcing the classification algorithm to concentrate on those observations that are more difficult to classify. Each iteration therefore produces a new classification tree, with the intent of correcting misclassification errors committed in the previous iteration.

The boosting algorithm implemented in C5.0 is based upon AdaBoost.M1 [23], [5]. Following [23], w_x^t is defined to be the weight assigned to observation x at trial t . For $t = 1$, $w_x^1 = 1/N$ for all x , where N is the total number of observations in the training set. At each iteration, a classifier C^t is constructed using the assumption that for each x , w_x^t reflects the probability of occurrence for x . An error term, ϵ^t is calculated as the sum of the weights of the misclassified observations at each iteration. The system terminates if $\epsilon^t > 0.5$ or if $\epsilon^t = 0$ (i.e., if $> 50\%$ of the observations are misclassified or if C^t classifies all instances correctly). At each iteration, for each observation that C^t correctly classifies a new weight is estimated as

$$w_x^{t+1} = w_x^t \times \epsilon^t / (1 - \epsilon^t). \quad (6)$$

Conversely, if the observation was not correctly classified, w_x is unchanged. Note that at each iteration, w_x is normalized such that $\sum w_x = 1$.

The result of this procedure is that a new tree with different errors is estimated at each step. The final, boosted classifier is then estimated by voting, where the vote for classifier C^t is worth $\log(1/\beta^t)$ units, where $\beta^t = \epsilon^t / (1 - \epsilon^t)$. Studies conducted by machine learning researchers using a variety of nonremote sensing data sets have shown that boosting tends to reduce misclassification error rates by about 25% on average, and that the improvement in classification accuracy tends to stabilize by about ten iterations [23].

IV. METHODS

A. Analysis

The analyses performed for this work examine questions related to feature selection and the utility of boosting for land cover classification from remotely sensed data. Previous research has explored the utility of phenological metrics by examining the separability of classes in feature space using global data at 1° spatial resolution in association with maximum likelihood classification techniques [2]. More recently, phenological metrics have been tested in association with decision trees using the 8 km AVHRR pathfinder data set [3].

Here we consider similar questions using C5.0. We perform this analysis using the same data as that used in [2], and also using data at 1 km spatial resolution over North America. Phenological metrics considered include the annual minimum, maximum, amplitude, and mean of monthly NDVI values. For the 1° data, geographic position is encoded in coordinates of latitude (0–180°, with the South Pole as 0) and longitude (0–360°). For the 1 km data, geographic position is encoded using row and sample coordinates from images geo-rectified to a Lambert Azimuthal equal area projection. The inclusion of geographic position is based on the hypothesis that because large scale climate patterns exert strong control on the geographic distribution of vegetation biomes, geographic position serves as a good predictor of land cover and vegetation class at continental to global scales. As part of this analysis we assess the utility of boosting by comparing cross validated classification results derived from a single (unboosted) decision tree to those produced by boosted decision tree classifications estimated from the same training data.

B. Data

The paucity of high quality training data available for training and testing of classification algorithms at continental scales has precluded previous detailed studies of this nature and remains a limiting consideration. For this work, we have used two data sets. First, we used the North America IGBP classification map and associated twelve month time series of AVHRR NDVI data produced by EROS Data Center (EDC) [14]. These data provide IGBP land cover labels at 1 km spatial resolution for the entire North American Continent. It is important to note that although finite levels of labeling error are present, the map does represent the best of its kind for North America. It was generated by manual procedures involving unsupervised clustering of maximum value NDVI observations composited over North America at monthly time steps for the period from April of 1992 to March of 1993. Land cover labels were assigned to each 1 km pixel by manual splitting and labeling of NDVI clusters using extensive ancillary data related to soils, climate, topography and other relevant information.

Ideally, we would prefer to use training and test data derived from *in-situ* observations, aerial photography, or even manually classified Landsat data. Indeed, efforts are currently underway to compile a global database of site data that will be used to produce and assess the land cover maps produced from MODIS data. Unfortunately, these data are not yet available. We have therefore relied on the IGBP map produced at EDC under the assumption that it represents the best available source of this type of data. Using the EDC IGBP database, a random sample of 5545 joint observations of NDVI data and associated IGBP class values were extracted and used for the analyses presented here. The IGBP classification and the frequency distribution of IGBP classes within the random sample is presented in Table I.

The second data set we examine is composed of a time series of AVHRR NDVI measurements collected at monthly time intervals during 1987. These data were extracted from

TABLE II
NDVI-1 DEGREE GLOBAL LAND COVER CLASSES
(N = NUMBER OF SAMPLES IN EACH CLASS)

	Class Name	N
1	broadleaf evergreen forest	628
2	coniferous forest & woodland	320
3	high latitude deciduous forest & woodland	112
4	tundra	735
5	deciduous-evergreen forest & woodland	57
6	wooded grassland	212
7	grassland	348
8	bare ground	291
9	cultivated	527
10	broadleaf deciduous forest & woodland	15
11	shrubs and bare ground	153

a global dataset compiled as part of the International Land Surface Climatology Project (ISLSCP) Initiative I CDROM [17], and include one maximum value NDVI composite value for each month of 1987. For details, the reader is referred to [13]. The specific training data and associated class labels were compiled by DeFries and Townshend [4]. These observations and labels include 3398 1° × 1° locations where three widely used maps of land cover and vegetation [16], [20], [31] are in agreement. The classification scheme used to label these data and their associated class frequency distribution is presented in Table II.

It is important to note that both sets of NDVI data include finite levels of noise associated with the process used to composite the data at monthly time steps [10]. A variety of work has examined these issues [8], [13], [30]. In particular, Myneni [19] provides a systematic analysis of the combined effects of the atmosphere and surface bidirectional reflectance on NDVI measurements collected from satellites. Further, cloud contamination, large solar zenith angles, bias in view zenith angles, and registration errors in the monthly AVHRR composite data all contribute noise to these data [32]. For the purposes of classification using decision trees, the key issue is whether or not the noise is systematic and of sufficient magnitude to cause confusion between classes. For this work we assume that this is not the case based on the success of previous work in classifying land cover from monthly composites of AVHRR data (e.g., [4]). Complete details regarding the processing techniques used to generate each of the data sets used here are provided in [4], [13], [32], [14].

V. RESULTS

The questions examined in this paper are addressed by comparing classification accuracies achieved using different feature sets and boosted versus unboosted decision trees. To provide the most realistic and robust estimates of classifier performance, we performed a ten-fold cross validation for each classification case considered. To do this, the data were randomly partitioned into ten equal sized subsets, ensuring that the class distribution of the entire dataset was maintained in each. For each run, one subset was held out, using the nine remaining subsets for training and pruning. The reserved subset was then used to estimate the predicted classification accuracy of the decision tree for unseen data, thereby ensuring that our training

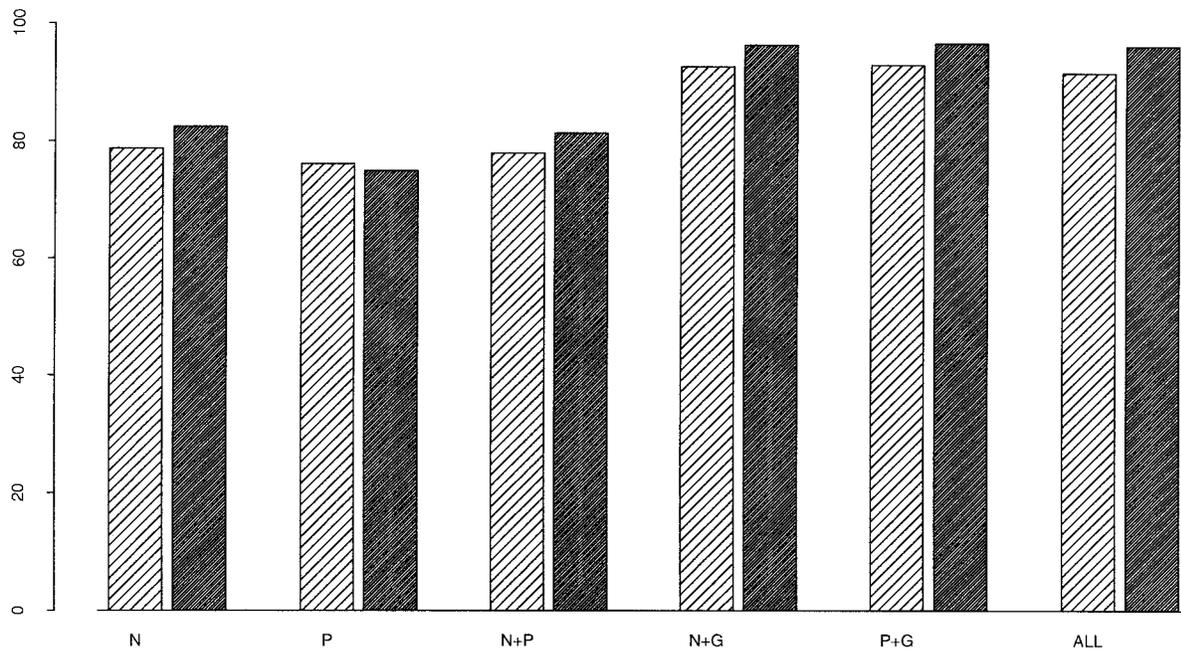


Fig. 2. Cross validated classification accuracies for decision trees estimated using different input features for the 1° global data set. Results for unboosted and boosted trees are plotted in light and dark shades, respectively. (N = NDVI alone; P = phenological metrics alone; N + P = NDVI and phenologic metrics; N + G = NDVI and geographic position; P + G = phenologic metrics and geographic position; ALL = all input features used.)

and testing data sets were independent for each run. Six input feature data sets were generated using different combinations of input features: NDVI only, phenologic metrics only, NDVI and phenological metrics, NDVI and geographic position, phenologic metrics and geographical position, and NDVI and phenological metrics and geographic position. Classification trees were generated with and without boosting using C5.0. Values for classification accuracies presented below represent average values across the ten cross validation runs. Results from each of our classification exercises are summarized in Table III.

A. Boosting

Classification accuracies for decision trees estimated from the 1° global data are shown in Fig. 2, and for the 1 km North America data in Fig. 3. Cross validated classification accuracies from unboosted trees are plotted in the lighter shaded bars and accuracies produced by boosted trees are plotted in the darker shaded bars. These results show that boosting improves classification accuracies for most of the cases tested. Three exceptions to this pattern are noted. Specifically, for the feature set composed of phenologic metrics alone (both 1 km and 1° data) and the feature set composed of all possible features (1 km data only), boosting resulted in slightly lower accuracies.

It is interesting to note that the improvement yielded by boosting is not consistent between the 1° and 1 km data sets. For the 1 km data, improvements were generally on the order of 7–9% (excluding the case composed of all features). For the 1° data, on the other hand, boosting tended to improve classification accuracy by about 4%. However, it is important to note these improvements in classification accuracy account

TABLE III
SUMMARY OF RESULTS FROM BOOSTED AND UNBOOSTED DECISION TREES.
NOTE: NO VALUES (*) ARE PROVIDED FOR THE NUMBER OF NODES FOR BOOSTED CLASSIFICATION BECAUSE THESE CLASSIFICATIONS ARE ESTIMATED FROM MULTIPLE DECISION TREES WITH DIFFERING NUMBERS OF NODES

Input Features	Accuracy (%)	# of Nodes
1 Degree: NDVI Only	78.7	221.5
1 Degree: NDVI Only, boosted	82.4	*
1 Degree: Phenology Only	76.1	134.7
1 Degree: Phenology Only, boosted	74.9	*
1 Degree: NDVI+Phenology	77.9	226.5
1 Degree: NDVI+Phenology, boosted	81.3	*
1 Degree: NDVI+Position	92.6	112.4
1 Degree: NDVI+Position, boosted	96.3	*
1 Degree: Phenology + Position	94.9	103.0
1 Degree: Phenology + Position, boosted	96.6	*
1 Degree: All	91.6	111.5
1 Degree: All, boosted	96.1	*
1 km: NDVI Only	67.4	548.3
1 km: NDVI Only, Boosted	76.3	*
1 km: Phenology Only	56.7	461.0
1 km: Phenology Only, boosted	54.9	*
1 km: NDVI+Phenology	67.0	569.9
1 km: NDVI+Phenology, boosted	75.4	*
1 km: NDVI+Position	72.4	325.2
1 km: NDVI+Position, boosted	79.5	*
1 km: Phenology + Position	62.8	568.0
1 km: Phenology + Position, boosted	66.2	*
1 km: All	77.8	345.9
1 km: All, boosted	76.2	*

for anywhere from ≈ 20 –50% of the misclassified samples (i.e., ≈ 20 –50% of misclassified training observations from unboosted decision trees are correctly classified by the boosted trees). Further, the overall improvement gained from using the best feature set in association with boosting improved classification accuracies from 78.7–96.6% (1° data) and from

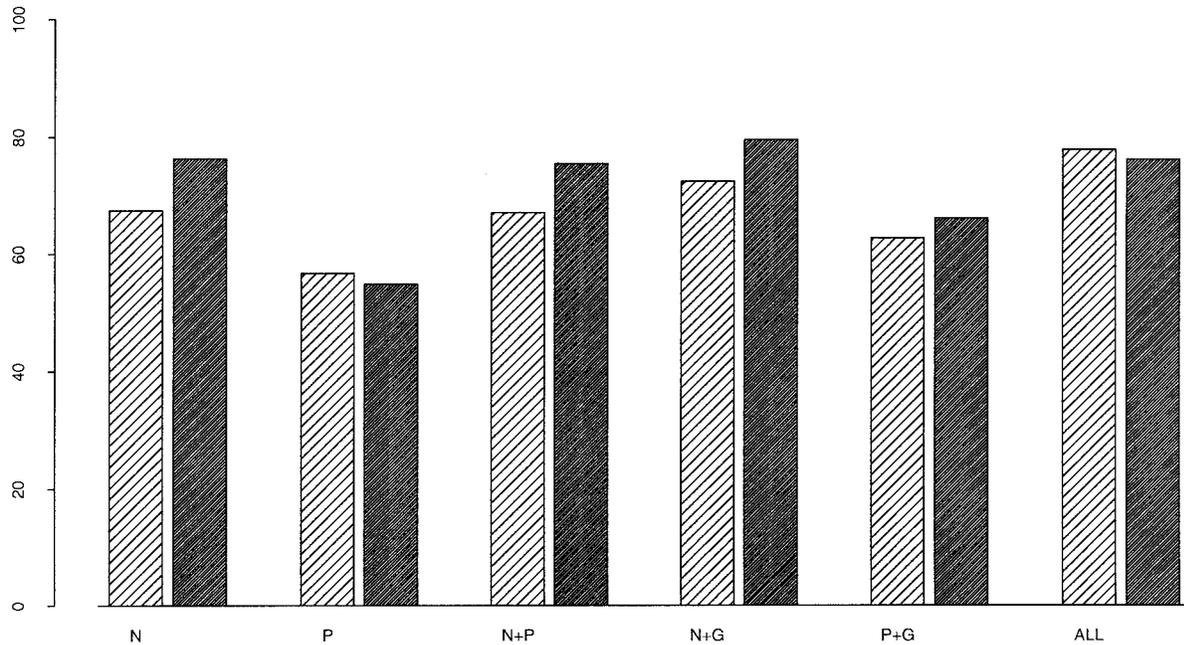


Fig. 3. Same as Fig. 2, but for the 1 km data for North America.

67.4–79.5% (1 km data) relative to unboosted classifications estimated from NDVI data alone. Stated another way, the use of all available input features in association with boosting reduced misclassification rates by 84 and 37% for the global and North America data sets, respectively, relative to unboosted trees estimated from NDVI data only.

The boosted decision tree classifications were estimated using ten iterations of the decision tree algorithm. We choose to use ten iterations because previous studies using nonremote sensing data sets suggest that this number of iterations provides maximum improvement in classification accuracy and that little is gained by performing additional boosting runs [5]. To test this guideline, we estimated boosted classification trees using the full feature space for both data sets and varied the number of boosting iterations from two to 15. Results from this analysis are presented in Fig. 4 (Note the use of different scales on the Y-axes). These results confirm that the accuracy improvement achieved through boosting approaches an asymptotic value after a relatively few number of iterations. Indeed, the results presented here suggest that relatively little accuracy is gained beyond about seven iterations.

B. Feature Selection

Patterns in classification accuracy among the different feature sets were generally consistent between unboosted and boosted trees, but clear differences are observed between boosted and unboosted results for each feature set. For the 1° global data set, decision trees estimated using different combinations of input features not including geographic position produced comparable accuracies. In comparison, classification trees estimated from feature sets that include geographic location show considerably (≈ 14 –18%) higher accuracies. Classifications estimated using only phenological metrics produced comparable accuracies to those estimated from the full

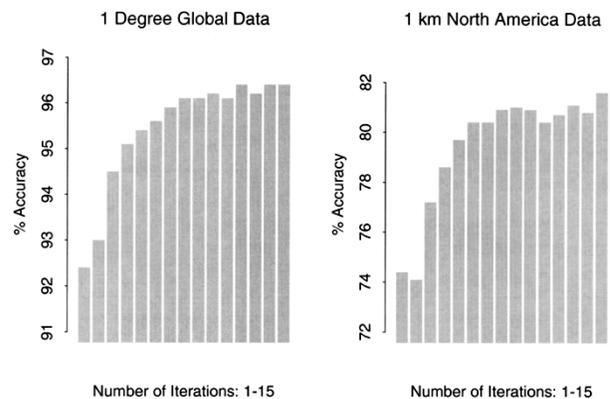


Fig. 4. Cross validated classification accuracies for boosted decision trees for varying numbers of boosting iterations.

12 month time series, but combining these two feature sets yielded no improvement in classification accuracy.

For the 1 km data, differences in classification results produced among the different input features are more subtle. In contrast to the 1° global data, classification trees estimated from phenological metrics produced accuracies that were lower than those estimated from the full twelve month time series. Further, the use of geographic position provided less improvement relative to those achieved in the 1° global data.

Overall, the highest classification accuracies were produced using a combination of phenology and position for the 1° data, and a combination of the original twelve month NDVI data set and geographic position for the 1 km data set. The fact that phenology provides quite poor results for the 1 km data suggests that subtle information in the twelve month data is not present in the phenological metrics and is required to provide the highest accuracy. Also, note that a by-product of improved classification accuracy produced

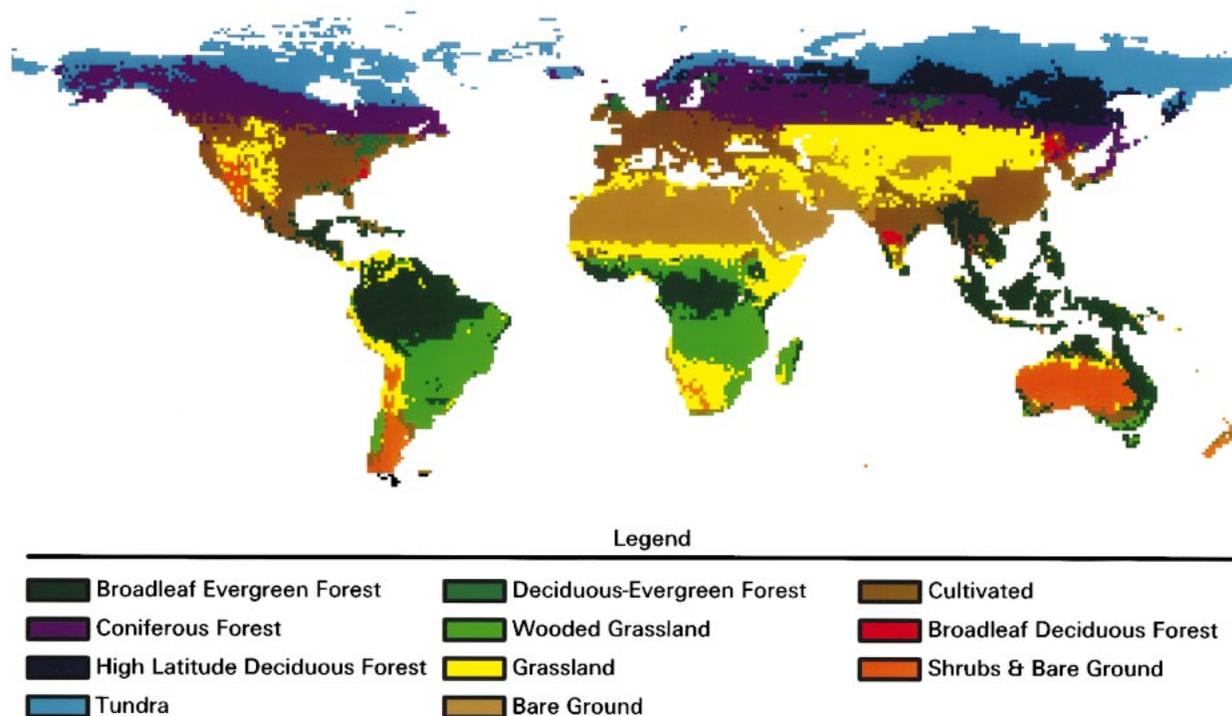


Fig. 5. Map of global vegetation produced from the decision tree estimated using all input features at 1° spatial resolution.

by including geographic position as an input feature is that tree complexity tends to decrease with classification accuracy (Table III). Indeed, the number of nodes in a tree is a good indicator of the predictive power of the input features provided to a decision tree estimation algorithm. Therefore, in addition to being more accurate, decision trees estimated using features with high discrimination among the classes are more compact and accurate than trees estimated from features with less predictive power.

VI. CONCLUSION

The general objective of the work described here is to assess two strategies designed to maximize land cover classification accuracies derived from supervised classification algorithms being developed for use with MODIS data. The specific objectives are to improve our understanding of how supervised classification algorithms interact with training data, and what the impact of these interactions is on the final map produced by classification of coarse spatial resolution remote sensing data. Because MODIS will provide data that are superior to AVHRR in terms of radiometric quality, geometric integrity, and spectral resolution, we expect accuracies derived from classifications based on MODIS data to improve accordingly. However, improved understanding of the utility of currently available input features as well as careful accounting for artifacts introduced by training site selection are required to provide the best product possible.

The results presented in this paper suggest several main conclusions. First, boosting improved the classification accuracy for nine of the twelve input feature data set combinations examined. The improvement was substantial in many cases,

and dramatic in some. We therefore conclude that boosting is a useful technique and should be used for land cover classification problems using remotely sensed data at continental to global scales.

Second, adding features related to vegetation phenology produced little improvement to classification accuracy. This result is somewhat at odds with the conclusions of DeFries *et al.* [2] who found that phenological metrics provide useful information to classifications performed using data compiled at global scales. DeFries *et al.* [2] used maximum likelihood techniques, however, which are better suited for use with summarizing variables such as phenological metrics (which tend to be more Gaussian and less noisy than the NDVI data from which they are derived). More recent work using the decision tree classification algorithm in Splus (based upon the CART model [1]) also supports the use of phenological metrics [3]. The likely explanation for the apparent contradiction between the results cited in [2], [3] and those presented here is that developments in decision tree algorithms since CART have produced algorithms that are superior in terms of their ability to handle noise and perform feature selection. Indeed, the fact that the decision tree classification accuracies presented in this work show no improvement with the addition of phenological metrics suggests that the useful information provided in the phenological metrics is being extracted by the trees from the original NDVI data.

Third, for the data sets examined here, the use of geographic position provides substantial predictive power to the decision tree classification algorithms. As we indicated in Section IV-A, this result can be largely explained in terms of climate control on the large scale distribution of global vegetation.

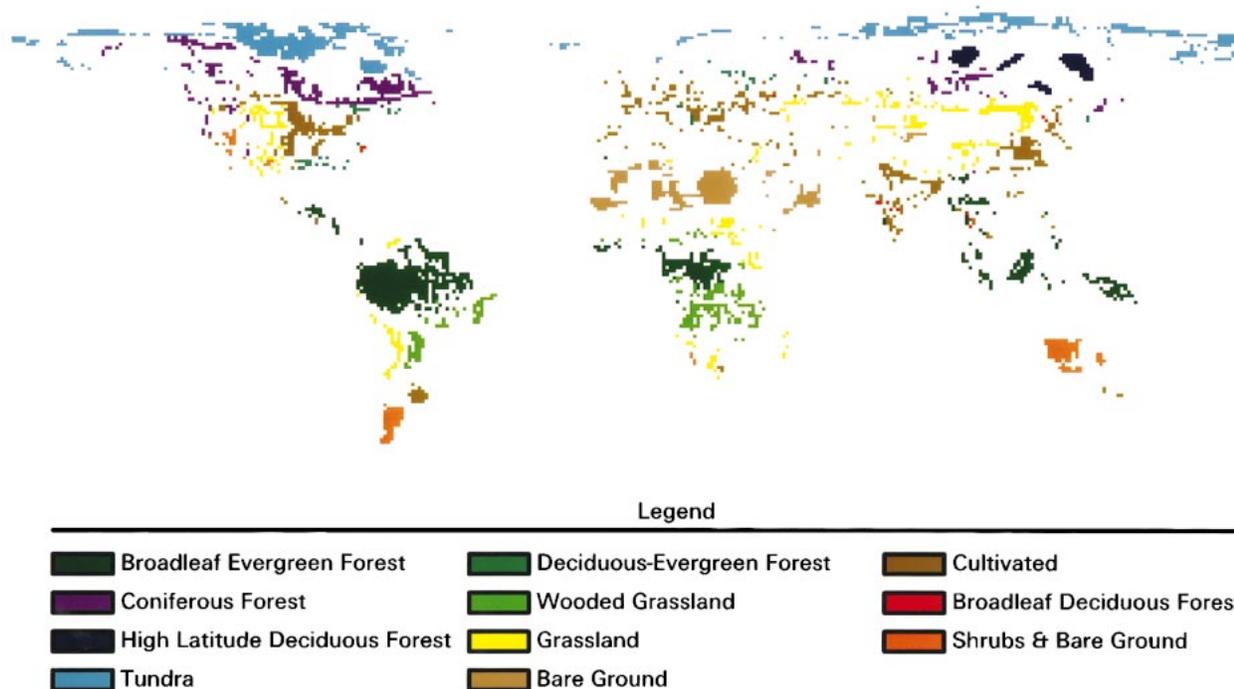


Fig. 6. Map of training sites used to produce the decision tree at 1° spatial resolution.

Stated another way, it is not surprising that geographic position has relatively high predictive power when classifying fairly coarse classes of vegetation at continental and global scales. This effect is particularly evident in the 1° data for which geographic position increases classification accuracies by more than 13%.

Despite this encouraging result, a word of caution is in order. To illustrate, Fig. 5 presents a map of global vegetation generated using the decision tree estimated from all available features using the 1° data set. The cross validated classification accuracy for the training data for this input feature set was 96.1%. However, visual inspection of this map shows distinct latitudinal banding in eastern North America and Eurasia at roughly 50° north that is clearly a by-product of interaction between the geographic location of the training data and the classification procedure. In North America, evergreen coniferous forests are almost completely absent in the western mountain regions of the United States (replaced by grassland and agriculture classes), and deciduous and conifer forests of the southern and eastern United States have been labeled as cultivated. Further inspection reveals a variety of other problems.

These observations clearly show that the cross validated estimate of classification accuracy for these data is spurious. In particular, this result seems to be produced by interaction between the decision tree estimation algorithm and the distribution of the training data sites (Fig. 6). Because the decision tree attempts to optimize classification accuracy with respect to the training data provided, over- or underrepresentation of specific classes within geographic subregions introduces substantial bias to classifications using geographic position as

an input feature. Stated another way, because the training data are not distributed evenly within the geographic space of each class, a classification based partly on geographic coordinates proves to be very effective for classifying the training data. Unfortunately, these accuracies do not reflect the true accuracy of the final map produced from the decision trees estimated from these data.

In contrast, the geographic distribution of sample points in the 1 km North America data is random (i.e., evenly distributed geographically) and the improvement in classification accuracies yielded by inclusion of geographic position is substantially smaller relative to that produced for the 1° global data set. Therefore, the improvement in classification accuracy achieved by inclusion of geographic features in this data set is probably more representative of the true predictive utility of these features. An important conclusion from these results is therefore that geographic position should only be used as a secondary input feature used to discriminate between land cover classes that are spectrally similar, but geographically distinct.

From a more general perspective, it is clear that the classification results produced by supervised algorithms are heavily reliant on the quality and representativeness of the training data used. Thus, care must be used in interpreting estimated classification accuracies from remote sensing derived maps at continental to global scales, and that by extension, geographically and spectrally representative training data are a key requirement to the success of supervised classification algorithms planned for use with MODIS data. Indeed, probably the most important factor influencing the quality of land cover maps produced from MODIS data will be the quality of the

training data used. To this end, the compilation of extensive and high quality training data are a current focus of our efforts.

ACKNOWLEDGMENT

The authors would like to thank R. DeFries for kindly supplying the 1° data, the Staff of the Global Land 1-KM AVHRR Project, Eros Data Center, for supplying the 1 Km data for North America, and J. Hodges for generating the color graphics.

REFERENCES

- [1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
- [2] R. DeFries, M. Hansen, and J. R. G. Townshend, "Global discrimination of land cover types from metrics derived from AVHRR pathfinder data," *Remote Sensing Environ.*, vol. 54, pp. 209–222, 1995.
- [3] R. S. DeFries, M. Hansen, J. G. R. 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," *Int. J. Remote Sensing*, vol. 19, no. 16, pp. 3141–3168, 1998.
- [4] R. S. DeFries and J. G. R. Townshend, "NDVI-derived land cover classifications at a global scale," *Int. J. Remote Sensing*, vol. 5, pp. 3567–3586, 1994.
- [5] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proc. Computat. Learning Theory: 2nd Euro. Conf. EuroCOLT'95*, 1996, vol., pp. 23–27.
- [6] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sensing Environ.*, vol. 61, pp. 399–409, 1997.
- [7] S. Goward, C. Tucker, and D. Dye, "North American vegetation patterns observed with the NOAA-7 advanced very high resolution radiometer," *Vegetatio*, vol. 15, pp. 237–253, 1985.
- [8] G. Gutman and A. Ignatov, "Global land monitoring from AVHRR: Potential and limitations," *Int. J. Remote Sensing*, vol. 16, pp. 2301–2309, 1995.
- [9] M. Hansen, R. Dubayah, and R. DeFries, "Classification trees: An alternative to traditional land cover classifiers," *Int. J. Remote Sensing*, vol. 17, pp. 1075–1081, 1996.
- [10] B. Holben, "Characteristics of maximum-value composite images from temporal AVHRR data," *Int. J. Remote Sensing*, vol. 5, pp. 145–160, 1986.
- [11] C. Justice, J. Townshend, B. Holben, and C. Tucker, "Analysis of the phenology of global vegetation using meteorological satellite data," *Int. J. Remote Sensing*, vol. 6, pp. 1271–1318, 1985.
- [12] D. Lloyd, "A phenological classification of terrestrial vegetation cover using shortwave vegetation index imagery," *Int. J. Remote Sensing*, vol. 11, pp. 2269–2279, 1990.
- [13] S. O. Los, C. O. Justice, and C. J. Tucker, "A global 1 degree by 1 degree NDVI data set for climate studies derived from the GIMMS continental NDVI data," *Int. J. Remote Sensing*, vol. 15, pp. 3493–3519, 1994.
- [14] T. R. Loveland and A. S. Belward, "The IGBP-DIS global 1 km land cover data set, DISCover: First results," *Int. J. Remote Sensing*, vol. 18, pp. 3289–3295, 1997.
- [15] T. R. Loveland, J. Merchant, J. Brown, D. O. Ohlen, B. C. Reed, P. Olson, and J. Hutchinson, "Seasonal land cover regions of the United States," *Annals Assoc. Amer. Geogr.*, vol. 85, pp. 339–355, 1995.
- [16] E. Matthews, "Global vegetation and land use: New high resolution databases for climate studies," *J. Climate Appl. Meteorol.*, vol. 22, pp. 474–487, 1986.
- [17] B. W. Meeson, F. E. Corprewand J. McManus, D. M. Myers, J. W. Closs, K. J. Sun, D. J. Sunday, and P. Sellers, "Isiscp initiative i-global data sets for land-atmosphere models, 1987-1988," *Volumes 1–5. Published on CD by NASA (USA_NASA_GDAAC_ISLSCP_001-USA_NASA_GDAAC_ISLSCP_005*, 1995.
- [18] J. Mingers, "An empirical comparison of pruning methods for decision tree induction," *Mach. Learn.*, vol. 4, pp. 227–243, 1989.
- [19] R. B. Myneni, "Atmospheric effects and spectral vegetation indices," *Remote Sensing Environ.*, vol. 47, pp. 390–402, 1994.
- [20] J. S. Olsen, J. Watts, and L. Allison, "Carbon in live vegetation of major world ecosystems." Tech. Rep. W-7405-ENG-26, U.S. Dept. of Energy, Oak Ridge Nat. Lab., CA, 1983.
- [21] J. R. Quinlan, "Simplifying decision trees," *Int. J. Man-Mach. Stud.*, vol. 27, pp. 221–234, 1987.
- [22] ———, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
- [23] ———, "Bagging, boosting, and c4.5," in *Proc. 13th Nat. Conf. Artificial Intell.*, Portland, OR, 1996, pp. 725–730.
- [24] S. W. Running, C. O. Justice, V. Salomonson, D. Hall, J. Barker, Y. J. Kaufmann, A. H. Strahler, A. R. Huete, J. P. Mullerand V. Vanderbilt, Z. M. Wan, P. Teillet, and D. Carneggie, "Terrestrial remote sensing science and algorithms planned for EOS/MODIS," *Int. J. Remote Sensing*, vol. 15, pp. 3587–3620, 1994.
- [25] S. R. Savafian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, pp. 660–674, 1991.
- [26] R. E. Shapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, 1990.
- [27] A. H. Strahler and J. Townshend, *MODIS Land Cover Product Algorithm Theoretical Basis Document (ATBD), V4.1*, Boston Univ. Ctr. Remote Sensing, Boston, MA, Dec. 1996.
- [28] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Trans. Geosci. Remote Sensing*, vol. GE-15, pp. 142–147, 1977.
- [29] J. Townshend and C. Justice, "Analysis of the dynamics of African vegetation using the normalized difference vegetation index," *Int. J. Remote Sensing*, vol. 7, pp. 1435–1445, 1986.
- [30] C. J. Tucker, "History of the use of AVHRR data for land applications," In G. D'Souza, A. S. Belward, and J. P. Malingreau, eds., *Advances in the Use of NOAA AVHRR Data for Land Applications*. Brussels, Belgium: Kluwer, 1996, pp. 1–19.
- [31] M. F. Wilson and A. Henderson-Sellers, "A global archive of land cover and soils data for use in general circulation models," *J. Climatol.*, vol. 5, pp. 119–143, 1985.
- [32] Z. L. Zhu and L. Yang, "Characteristics of the 1 km AVHRR data set for North America," *Int. J. Remote Sensing*, vol. 17, pp. 1915–1924, 1996.

Mark A. Friedl received the B.S. degree from McGill University, Montreal, P.Q., Canada, in 1986, and the Ph.D. degree from the University of California, Santa Barbara, in 1993.

He is an Assistant Professor at Boston University, Boston, MA. His research interests include remote sensing, geographical analysis and modeling using GIS, and biophysical modeling of land surfaces using remote sensing and GIS. Recently, his research has examined the use of machine learning tools in remote sensing problems including classification of multitemporal satellite data and automated detection of errors in training data for supervised classification algorithms.

Dr. Friedl is a member of the American Geophysical Union, the American Meteorological Society, and the American Association for the Advancement of Science.

Carla E. Brodley received the B.S. degree from McGill University, Montreal, P.Q., Canada, in 1985 and the Ph.D. degree in computer science from the University of Massachusetts, Boston, in 1994.

She is an Assistant Professor in the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN. Her research interests include machine learning, computer vision, pattern recognition, and knowledge discovery in databases. She has worked in the areas of anomaly detection, classifier formation, feature selection, and content-based image retrieval. She has applied techniques from these areas to problems from a variety of fields including remote sensing, medical images, and computer security.

Alan H. Strahler (M'86), for a photograph and biography, see this issue, p. 738.