ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Remote Sensing of Environment 90 (2004) 331–336

Remote Sensing of Environment

www.elsevier.com/locate/rse

Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis

Rick Lawrence[a,*], Andrew Bunn[a], Scott Powell[b], Michael Zambon[a]

[a] Department of Land Resources and Environmental Sciences, Montana State University, PO Box 173490, Bozeman, MT 59717, USA
[b] Ecology Department, Montana State University, Bozeman, MT 59717, USA

Received 11 September 2003; received in revised form 13 January 2004; accepted 14 January 2004

Abstract

Classification tree analysis (CTA) provides an effective suite of algorithms for classifying remotely sensed data, but it has the limitations of (1) not searching for optimal tree structures and (2) being adversely affected by outliers, inaccurate training data, and unbalanced data sets. Stochastic gradient boosting (SGB) is a refinement of standard CTA that attempts to minimize these limitations by (1) using classification errors to iteratively refine the trees using a random sample of the training data and (2) combining the multiple trees iteratively developed to classify the data. We compared traditional CTA results to SGB for three remote sensing based data sets, an IKONOS image from the Sierra Nevada Mountains of California, a Probe-1 hyperspectral image from the Virginia City mining district of Montana, and a series of Landsat ETM+ images from the Greater Yellowstone Ecosystem (GYE). SGB improved the overall accuracy of the IKONOS classification from 84% to 95% and the Probe-1 classification from 83% to 93%. The worst performing classes using CTA exhibited the largest increases in class accuracy using SGB. A slight decrease in overall classification accuracy resulted from the SGB analysis of the Landsat data.

Keywords: Classification tree analysis; Stochastic gradient boosting; Accuracy

1. Introduction

Classification tree analyses (CTA; also referred to as classification and regression trees [CART] or decision trees) are increasingly being used for analysis and classification of remotely sensed digital imagery. CTA has been used successfully for classification of multispectral (Friedl & Brodley, 1997; Hansen et al., 1996) and hyperspectral imagery (Lawrence & Labus, 2003), incorporation of ancillary data with multispectral imagery for increased classification accuracy (Lawrence & Wright, 2001), and change detection analysis (Rogan et al., 2003). Although CTA is a relatively new statistical technique, having been developed about 20 years ago (Breiman et al., 1984), it has subsequently been the subject of considerable development and refinement. We examined whether one of the most recent statistical techniques designed to improve on CTA, stochastic gradient boosting (SGB), offers substantial advantages over traditional CTA approaches being used by remote sensing analysts.

CTA typically operates by recursively parsing the training observations based on a binary splitting measure applied to explanatory variables, such as spectral responses (Breiman et al., 1984; Lawrence & Ripple, 2000). The results of CTA are often in the form of easily interpreted dichotomous trees that can be used as classification rules either by themselves in rule-based classifiers or combined with expert knowledge (Lawrence & Wright, 2001). The increasing popularity of CTA as a classification technique stems from several advantages of CTA over traditional methods, such as maximum likelihood classifiers. CTA does not rely on any assumptions regarding the distribution of the data, since, unlike some conventional classifiers, it is a non-parametric technique. A wide diversity of data sources can be used as inputs to the classification, such as raw spectral bands, derived spectral information (such as tasseled cap components or vegetation indices), topographic data, and other GIS layers. CTA automatically selects the best data layers for classification from those provided by the analyst. CTA handles continuous and categorical information equally well, while traditional classifiers cannot include categorical data. Most importantly, in many reported comparisons, CTA has resulted in higher accuracies than other methods [e.g.,

* Corresponding author. Tel.: +1-406-994-5409; fax: +1-406-994-5122.
E-mail address: rickl@montana.edu (R. Lawrence).

0034-4257/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2004.01.007

Lawrence & Labus, 2003; Pal and Mather, 2003 (but see results with higher dimensional data); Rogan et al., 2002].

As a statistical method for classifying data, CTA has several problems that have been noted (Friedman, 2001). First, CTA will not necessarily produce the optimal classification tree, since it partitions the data on a one-step-at-a-time basis. A split at one level that does not create the best classes at that level might enable better splits at lower levels and a better overall classification tree, much the way sacrificing a piece in a game of chess can result in a better position several moves ahead, but this is not provided for in the methodology. In addition, both inaccuracies and outliers in the training data can adversely affect CTA because such data can potentially account for a large portion of the variability in the data (Friedman, 2001). CTA algorithms can, therefore, concentrate on correctly classifying this erroneous or extreme data to the detriment of correctly classifying other data. This type of data is typical in remote sensing training data where, for example, a training polygon for rangeland containing 50 pixels might be expected to have several pixels with a predominance of bare ground. Finally, the presence of an unbalanced data set with some classes more heavily represented than others can affect the performance of CTA, with the analysis sometimes dividing heavily represented classes rather than splitting out lightly represented classes.

Methods for producing ''optimal'' classification trees have not yet been practical (Friedman, 2001). Several methods, however, including boosting and bagging, have recently been developed to address the shortcomings of CTA (Bauer & Kohavi, 1999; DeFries & Chan, 2000; Friedl et al., 1999). These methods, sometimes called voting or ensemble methods, operate by generating multiple trees and classifying observations generally based on a majority or weighted majority vote of the multiple trees (Opitz & Maclin, 1999). The primary difference among these new methods is how the multiple trees are developed. Two major types of methods have been developed, those that develop new classification trees based on the results of previous classification trees (boosting methods) and those that rely on subsets of the training data to develop new classification trees (bagging methods) (Bauer & Kohavi, 1999). Many variants of these basic methods exist (Opitz & Maclin, 1999).

Boosting methods have generally produced the greatest increases in accuracy, although under certain circumstances lower accuracies can result (Bauer & Kohavi, 1999; Opitz & Maclin, 1999). Boosting methods begin by producing a standard classification tree (Freund & Schapire, 1996, 1999). Training data are then assigned weights with incorrectly classified data given greater weight; the greater the misclassification, the greater the assigned weight. This forces the new classification tree to emphasize the hardest classification problems in the training data. The process is repeated for a specified number of iterations, and the set of resulting classification trees vote on the correct classification using a plurality rule. Boosting has been shown to improve classification tree performance in many cases, while performing at least as well as CTA in most remaining cases (e.g., Freund & Schapire, 1996, 1999; Opitz & Maclin, 1999). Boosting does not, however, assist with inaccurate training data, outliers, or unbalanced data sets. ''Outliers'' (training data that are incorrectly labeled or that are especially hard to distinguish from other classes), for example, can have an adverse effect on boosting because the algorithm will place emphasis on these observations, since they will be the worst classified and given the greatest weight in the boosted classification trees (Bauer & Kohavi, 1999; Freund & Schapire, 1996, 1999; Opitz & Maclin, 1999).

Bagging methods are bootstrapping approaches where multiple classification trees are developed by repeatedly selecting random subsets of the original training data (Breiman, 1996). A user-specified number of iterations is performed and, as in boosting, observations are classified based on the most common prediction from among the multiple classification trees. In a comparison of traditional classification trees, bagging, and boosting, bagging consistently produced higher classification accuracies than single classification trees, but was often less accurate than boosting (Opitz & Maclin, 1999).

SGB is a hybrid of the boosting and bagging approaches (Friedman, 2001, 2002). First, instead of using the entire data set to perform the boosting, a random sample of the data is selected at each step of the boosting process. Second, boosting is based on a steepest gradient algorithm, with the gradient defined by deviance (twice the binomial negative log-likelihood) as a surrogate for misclassification rates. Finally, instead of developing full classification trees at each stage of the boosting procedure, relatively small trees are developed, with 6 terminal nodes being a common size. As with the other ensemble methods, larger trees are not formed, rather each tree developed during the process (often 100–200 trees) is summed, and each observation is classified according to the most common classification among the trees. The combined effect of these differences from other boosting methods reduces SGBs sensitivity to inaccurate training data, outliers, and unbalanced data sets since, among other things, the steepest gradient algorithm places emphasis on misclassified training data that are close to their correct classification, rather than the worst classified data. SGB has been shown in most cases to produce substantially higher accuracies with independent data (data that were not used to develop the trees) than either CTA or other boosting methods (Friedman, 2002). Finally, unlike CTA, which is highly prone to overfitting to training data, SGB is highly resistant to overfitting since very small classification trees are used at each step of the boosting process.

2. Methods

We compared the accuracy of SGB to CTA on three different image classification problems. The three data sets were selected for their wide spectral, spatial, and land cover

Table 1
Comparative accuracies for classification of IKONOS imagery of the Sierra Nevada Mountains

	CTA accuracy (%)	SGB accuracy (%)
Producer's		
Tree	84	92
Water	96	96
Meadow	87	94
Rock	76	98
User's		
Tree	82	92
Water	100	100
Meadow	61	91
Rock	95	98
Overall	84	95

diversity and because we had used CTA previously to classify these data with varying degrees of success. Although these data represent a wide diversity, they are, of course, by no means exhaustive of all of the cases an image analyst might face.

The first data set included an IKONOS 4-m resolution multispectral image from Sequoia National Park, California acquired August 2001 (Bunn et al., in review). The IKONOS data consisted of four spectral bands from 450 to 850 nm. A slope gradient layer based on a 10-m digital elevation model (DEM) was resampled to 4 m based on nearest neighbor resampling and also included in the classification based on previous analysis. The classification was part of a study conducted to examine forest spatial patterns at upper treeline. The classification scheme, therefore, only included meadows, rock, trees, and water. The reference data collected through ground observations included 5560 sample points, which were randomly divided into equal training and accuracy assessment data sets.

The second data set was derived from a Probe-1 5-m resolution hyperspectral image of Virginia City, MT, acquired in August 1999 (Driscoll, 2002). Probe-1 is an across track sensor that collects spectral data in 128 bands from 440 to 2507 nm, all of which were used in the classifications. No ancillary data were incorporated into this classification. The area imaged has experienced substantial mineral extraction activities in the past, and the classification scheme included conifer and deciduous forest types, rangeland, water, disturbed lands from mining, and developed areas. The reference data collected through ground observations included 1947 sample points, which were randomly divided into approximately equal training and accuracy assessment data sets.

The final data set used was Landsat ETM+ imagery from the Greater Yellowstone Ecosystem (GYE) in Montana, Wyoming, and Idaho acquired in the summer, fall, and winter of 1999 and 2000 (registration error among all images less than 0.5 RMSE), together with extensive ancillary data (Lawrence & Wright, 2001). Six Landsat ETM+ spectral bands, ranging from 405 to 2350 nm, were used in the analysis (the thermal band was not used). Ancillary data that were used in the classifications included elevation, slope, and aspect from a 30-m DEM, tasseled cap brightness, greenness, and wetness components from each date, and difference images in these components between summer and fall, and summer and winter. The data were collected as part of a study of conifer forest expansion in the GYE, and the classification types included conifer and hardwood forests, herbaceous ground cover, conifer/herbaceous mix, conifer/hardwood mix, and burned areas. The reference data collected through air photo interpretation and ground observations included 1,122 sample points, which were randomly divided into equal training and accuracy assessment data sets.

We constructed classification trees for each of the training data sets using CART 5.0 (Salford Systems, 2002). CART 5.0 provides a variety of splitting rules for constructing classification trees, and in each case we selected the rule that maximized accuracy of the reserved accuracy assessment data (for the IKONOS and Landsat data, the class probability rule, and for the Probe-1 data, the gini rule). CTA often requires classification trees to be reduced, or pruned, to guard against overfitting to the training data. We used a cross validation method provided in CART 5.0 to select, in each case, the optimal size pruned tree. This pruning method randomly divides the original training data into 10 equal subsets for cross validation analysis. Classification tree sets, or groves, were created using SGB as implemented in TreeNet software (Salford Systems, 2001). The reserved accuracy assessment data sets were used to compare the accuracies of the CTA and SGB classifications using standard error matrix measures of accuracy.

3. Results

Summary accuracy statistics for the IKONOS data, including overall accuracy, producer's class accuracies, and user's class accuracies, demonstrated improved or equal

Table 2
Comparative accuracies for classification of Probe-1 hyperspectral imagery of Virginia City, MT, and surrounding areas

	CTA accuracy (%)	SGB accuracy (%)
Producer's		
Water	71	100
Conifer	96	91
Deciduous	50	88
Developed	74	96
Range	86	99
Disturbed	96	80
User's		
Water	100	99
Conifer	85	90
Deciduous	89	90
Developed	80	96
Range	94	94
Disturbed	88	89
Overall	83	93

accuracy using SGB compared to CTA (Table 1). Overall accuracy increased by 11%. SGB class accuracies improved for all measures except for producer's and user's accuracy for water, which stayed the same at 96% and 100%, respectively. In each case, the class accuracies that were the lowest using CTA realized the largest improvements in SGB, with producer's accuracy for rock increasing 22% from 76% to 98%, and user's accuracy for meadow increasing 30% from 61% to 91%. These improvements were the result of CTA exhibiting substantial confusion between these two classes, which was almost entirely resolved using SGB.

Summary statistics for the Probe-1 data exhibited similar results to the IKONOS data, but with some important differences (Table 2). Overall accuracy experienced a similar increase with SGB, in this case 10%. Also similar to the IKONOS data, the classes that experienced the lowest accuracies with CTA had the largest increases in accuracy using SGB, with producer's accuracy for deciduous increasing 38% from 50% to 88%, and user's accuracy for developed increasing 16% from 80% to 96%. Again, these improvements were almost entirely the result of confusion between these two classes using CTA that was almost entirely resolved using SGB. The accuracy of several classes, however, experienced decreases, contrary to the IKONOS classification.

Results for the Landsat classification were substantially different from the other data. Overall accuracy was nearly identical, decreasing 2% with SGB compared to CTA. The similarity in overall accuracy, however, was not shared by the class accuracies, which varied widely from CTA to SGB. Also contrary to the other data, no obvious trends in class accuracies were evident. For producer's accuracy, the worst performing class for CTA, conifer/herbaceous at 49%, was also the worst performing class with SGB, decreasing to 42% (Table 3). On the other hand, the next worse performing class with CTA, hardwood at 59%, had the largest accuracy increase with SGB, up 31%. The classes that performed well with CTA, conifer and conifer/hardwood, had substantial decreases in SGB producer's accuracy, down 11% and 29%, respectively. A similar lack of pattern was seen in the class user's accuracy statistics.

Table 3
Comparative accuracies for classification of Landsat ETM+ imagery of a portion of the Greater Yellowstone Ecosystem

	CTA accuracy (%)	SGB accuracy (%)
Producer's		
Conifer	90	79
Conifer/herbaceous	49	42
Burned	68	69
Conifer/hardwood	88	59
Hardwood	59	90
Herbaceous	83	96
User's		
Conifer	59	60
Conifer/herbaceous	88	85
Burned	67	28
Conifer/hardwood	32	36
Hardwood	89	74
Herbaceous	48	61
Overall	64	62

4. Discussion

Our results demonstrated that SGB can achieve substantially improved accuracy compared to CTA, although not in all cases. Although the CTA results for the IKONOS and Probe-1 data were very good in the low to middle 80% overall accuracy range, respectively, the results with SGB were outstanding in the middle 90% range. SGB was able to identify areas of class confusion in these data and resolve the discrepancies, generally without sacrificing the accuracy of other classes.

It is not clear why SGB failed to produce higher accuracies with the Landsat data and, in many instances, had reduced individual class accuracies. Analysis is problematic, since statisticians developing boosting methods remain unclear as to exactly why ensemble procedures produce superior results to single classification trees (Schapire et al., 1998), and our three data sets provide a limited sample from which to draw conclusions. We did, however, note certain patterns that lead us to working hypotheses.

The Probe-1 and IKONOS images represented high-resolution data, while the Landsat pixels are substantially coarser, covering from 36 to 56 times the area per pixel, respectively. Although the effectiveness of other boosting methods has been shown not to be dependent on high variance among classes (Schapire et al., 1998), high within class variability might have a positive effect on SGB compared to CTA. CTA might be unable to generate enough rules to cover all the possible variability present within each class, while the linear combination of trees in SGB might allow for many more ways to classify each observation (potentially the total number of different ways 100–200 trees could vote for each possible outcome). Higher resolution data generally exhibits greater within class variability than coarser resolution data, given similar class definitions. Further, the classification scheme for the Landsat data was more detailed (containing 6 vegetation classes versus 2 for the IKONOS data and 3 for the Probe-1 data), which also might have led to decreased within class variability. SGB, which developed between 150 and 200 decision trees to classify the data and combined these trees for prediction, was able to define a greater number of rules leading to specific classes. We hypothesized, therefore, that SGB might be more able to define a variety of rules to account for the increased class variability of the higher resolution data, thus resulting in greater increases in class accuracies. This is also consistent with previous findings that variance reduction by using random subsets of the data for boosting is an important reason why SGB outperforms other boosting methods (Friedman, 2002). In addition, other boosting methods have failed when the base classification tree had low accuracy (Schapire, 1999), and we might be seeing this effect with SGB and the low classification accuracy levels we achieved with the Landsat data.

As SGB conducts its boosting operation, it concentrates on resolving observations that are near the decision space boundaries defined by the data model (Friedman, 2002). That is, observations within an individual decision tree that are close to being in another class are more likely to be identified and corrected in the boosting operation. When two classes are very similar, therefore, SGB is particularly effective at resolving the differences between the two classes. This theoretical advantage in SGB was demonstrated in both the IKONOS and Probe-1 data. With the IKONOS data, the main source of confusion in the CTA classification was between rock and the alpine meadows found in the high Sierra Nevada Mountains in summer, where some areas can have spectrally similar dry, sparse grasses or rock on similar slopes. The main difference between the IKONOS CTA and SGB classifications was the resolution of these classes. Similarly with the Probe-1 data, the primary source of confusion with CTA was between the deciduous class and the developed class, which consisted of Virginia City, MT, a town containing substantial deciduous plantings. Again, SGB was able to resolve these similar classes.

Although SGB did not produce higher overall accuracy with the Landsat data, certain classes had substantial improvements. An analyst deciding between CTA and SGB might consider examining the results of the accuracy assessments to determine which classes to classify with CTA and which to classify with SGB.

An important disadvantage of SGB for some applications is that it does not provide readily interpretable decision trees. Such trees, which are provided by CTA, have numerous advantages. They reveal to the image analyst the basic structure of the data, identifying which variables (spectral bands and/or ancillary data) are being used to discriminate among classes. This information can be used to increase understanding of spectral properties of earth objects and their differences and can provide an important diagnostic on the adequacy of the classification model that cannot be provided by accuracy assessments alone. If the training and accuracy assessment data are not representative, for example, the classification model can develop erroneous rules that will not be revealed in the accuracy assessment. An examination of the rules can assist in this evaluation (e.g., Lawrence & Wright, 2001). This is not possible, however, with SGB, which develops multiple decision trees that are combined for prediction.

Although SGB does not provide readily interpretable decision trees or rules, it does provide information on the relative importance of variables in predicting each class. For the IKONOS data, for example, for predicting the tree class, blue was the most important band (relative importance of 100 assigned to the most important band), followed by slope (66), near infrared (43), green (32), and red (12). Considering interactions among variables, the most important pairs of variables for predicting the tree class were slope and blue, blue and near infrared, and slope and near infrared. These are logical rules and readily interpretable given the spectral response of the desired classes. The study site included many slope facets too steep to hold vegetation and therefore the slope layer was the obvious base layer for classifying the image. In areas with moderate to gentle slopes, the distinguishing spectral feature between the exposed substrate and the other classes was high blue reflectance in the granite, diorite, and monzonite spectra. In areas with slopes amenable to vegetation, the near-infrared layer distinguished well between the high chlorophyll wet meadows and the relatively lower chlorophyll forests. These class importance data provided some insight into the structure of the data, although not nearly as complete as that provided by CTA, where individual nodes can be traced to provide the precise rules applicable for each class.

SGB has the capability of producing higher accuracies than traditional CTA for remotely sensed data, although the results appear to be dependent on the specific data. As continues to be the case with image classification, there is no single classification algorithm that can be expected to provide maximum accuracies with all data since the statistical method that can best distinguish between and among classes is likely dependent upon the specific attributes of the data, including, among other things, classes being distinguished, resolution (spectral, spatial, radiometric, and temporal), and quality of training data. With that caveat in mind, SGB can provide truly exceptional accuracies with certain data. Substantially more experience is necessary to identify the best applications for SGB, although with our data the best performance was experienced with high-spatial-resolution imagery, and we speculate that this might be the result of high class variability present with such data. For the moment, it appears that SGB is a worthwhile alternative classification algorithm for an image analyst to consider.

References

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, *36*, 105–142.

Breiman, L. (1996). Bagging procedures. *Machine Learning*, *24*, 123–140.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Bunn, A. G., Waggoner, L. A., & Graumlich, L. J. (in review). Topographic mediation of growth of subalpine trees in the Sierra Nevada, USA. *Landscape Ecology*.

DeFries, R. S., & Chan, J. C. W. (2000). Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote Sensing of Environment*, *37*, 35–46.

Driscoll, S. G. (2002). *Detecting and mapping leafy spurge* (*Euphorbia esula*) *and spotted knapweed* (*Centaurea maculosa*) *in rangeland ecosystems using airborne digital imagery*. Bozeman, MT: Masters thesis, Montana State University.

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. *Machine learning: Proceedings of the Thirteenth International Conference* (pp. 148–156). San Francisco: Morgan Kaufman.

Freund, Y., & Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, *14*, 771–780.

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, *61*, 399–409.

Friedl, M. A., Brodley, C. E., & Strahler, A. H. (1999). Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *IEEE Transactions on Geoscience and Remote Sensing*, *GE-37*, 969–977.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.

Friedman, J. (2002). Stochastic gradient boosting: Nonlinear methods and data mining. *Computational Statistics and Data Analysis*, *38*, 367–378.

Hansen, M., Dubayah, R., & Defries, R. (1996). Classification trees: An alternative to traditional landcover classifiers. *International Journal of Remote Sensing*, *17*, 1075–1081.

Lawrence, R., & Labus, M. (2003). Early detection of douglas-fir beetle infestation with subcanopy resolution hyperspectral imagery. *Western Journal of Applied Forestry*, *18*, 202–206.

Lawrence, R. L., & Ripple, W. J. (2000). Fifteen years of revegetation of Mount St. Helens: A landscape-scale analysis. *Ecology*, *81*, 2742–2752.

Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering and Remote Sensing*, *67*, 1137–1142.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169–198.

Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, *86*, 554–565.

Rogan, J., Franklin, J., & Roberts, D. A. (2002). A comparison of methods for monitoring multitemporal vegetation change using Thematic Mapper imagery. *Remote Sensing of Environment*, *80*, 143–156.

Rogan, J., Miller, J., Stow, D., Franklin, J., Levien, L., & Fisher, C. (2003). Land-cover change monitoring with classification trees using Landsat TM and ancillary data. *Photogrammetric Engineering and Remote Sensing*, *69*, 793–804.

Salford Systems (2001). *TreeNet stochastic gradient boosting: An implementation of the MART methodology*. San Diego: Salford Systems.

Salford Systems (2002). *CART user's guide: An implementation of the original CART methodology*. San Diego: Salford Systems.

Schapire, R. E. (1999, December). Theoretical views of boosting and applications. *Proceedings of algorithmic learning theory, 10th International Conference, Tokyo, Japan.*

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, *26*, 1651–1687.