Land Cover Mapping Using Combination Classifiers

BRIAN M. STEELE[1]

[1] Dept. of Mathematical Sciences, The University of Montana, Missoula MT 59812; e-mail: bsteele@selway.umt.edu; telephone: 406-243-5396; fax: 406-243-2674.

Correspondence should be directed to the first author.

Abstract

In recent years, large scale land cover maps constructed from remotely sensed data have become important information sources for resource management. For many applications, poor map accuracy limits their usability. This article investigates methods of combining classification rules for improving accuracy in general, and for exploiting spatial information in particular. We examine the performance of these methods and the stacked regression method of Breiman (1996) and Mojirsheibani (1999) along with several variants. In our applications, a land cover map is a partition of an area into contiguous unclassified polygons which are assigned land cover types via a classification rule. Because polygons tend to differ with respect to land cover type, spatial association patterns are largely absent from polygon maps. However, there is some spatial information carried by the training observations. We propose a spatial classifier that uses spatially-close training observations for classification. While the spatial classifer is not particularly accurate, remarkable improvements in estimated accuracy were obtained when it was combined with linear discriminant and k-nearest neighbor classifiers.

Key Words: cross-validation, discriminant analysis, stacked regression, Remote sensing

1. Introduction

Geographic information systems (GIS) are widely used for large-scale resource analysis, monitoring, and management. An important component of many GIS is a land cover map showing the observed, or predicted, vegetation or land surface types. This article concerns land cover maps that are constructed by assigning land cover types to map polygons, or regions, using a classifier. Data are obtained from two sources. Spectral reflectance measurements are collected for all polygons by a remote sensing device such as the Landsat Thematic Mapper satellite. In addition, a training sample is collected by ground visitation of a subset of all map polygons. Then, a classification rule is constructed from the training sample and used with the remotely sensed data to predict land cover type for the unsampled polygons. While this approach can inexpensively map very large areas at a fine scale, classification, and hence, map accuracy is sometimes unsatisfactory for complex or wild landscapes such as those in the Rocky Mountains. Scott et al. (1993) and Moisen and Edwards (1999) provide examples of land cover map applications.

This article addresses several aspects of classification for land cover mapping. The principal topic is that of improving classification accuracy by combining multiple classifiers as a single classifier. Stacked regression methods (Breiman 1996; LeBlanc and Tibshirani 1996) are emphasized. I propose a very simple combination method which compares favorably with competing methods. The performance of the combination methods is compared by applying them to several training sets.

2. LAND COVER MAPS

A land cover map consists of a set of contiguous and disjoint polygons which are labeled according to the observed, or predicted, dominant vegetation or surface type (e.g., lodgepole pine dominated forest or xeric shrubland). In digital form, a land cover map amounts to a set of r-tuples denoting the observed, or predicted, land cover type and other variables (e.g., elevation) for each map polygon. One approach to land cover mapping constructs a base map of unclassified polygons from a Landsat Thematic Mapper (TM) satellite scene. Then, polygons

are assigned land cover type using a classifier (Foody, McCulloch and Yates 1995; Homer, Ramsey, Edwards, and Falconer 1997). A TM scene consists of a lattice of approximately 37.8×10^6 pixels of dimension 30×30 m; for each pixel, and spectral reflectance intensity is recorded for eight bands of the electromagnetic spectrum. The total area covered by a TM scene is approximately $34{,}000$ km^2 (Sabins 1997). Figure 1 shows two TM scenes, training sample locations and topographic relief.

Usually, polygons are constructed by aggregrating adjacent and spectrally similar pixels using image segmentation algorithms (Homer et al. 1997; Ryherd and Woodcock 1996; Woodcock and Harward 1992). Polygons are preferred as mapping units because they tend to be relatively homogeneous with respect to land cover and much more easily located for ground sampling. Computational effort of map construction and use is greatly reduced because the number of polygons is substantially fewer than the number of pixels. For example, the two TM scenes discussed in this article consist of 703701 and 813246 polygons, respectively. Polygons vary in size and shape, and adjacent polygons often differ with respect to land cover type. Consequently, spatial correlation among neighboring polygons tends to be weak and variable.

3. CLASSIFIERS

Suppose that a training sample $\boldsymbol{x} = \{ x_1, \ldots, x_n \}$ has been collected by random sampling of a population $\mathcal{P}$ consisting of g subpopulations, or groups $G_1, \ldots, G_g$. The ith observation in $\boldsymbol{x}$ is a triple denoted by $x_i = (t_i, y_i, z_i)$ where t_i is a covariate vector, y_i is a group label identifying group membership, and z_i is a location coordinate. For the remaining unclassified pairs in $\mathcal{P}$, a covariate vector t_0 and location z_0, are observed, but the group label y_0 is unobserved. The posterior probability that t_0 belongs to G_h is denoted by $P[y_0 = h \mid t_0]$. A general approach to combining classifiers can be formulated by treating classifiers as estimators of $(P[y_0 = 1 \mid t_0], \ldots, P[y_0 = g \mid t_0])$ which assign x_0 to the group with the largest posterior probability estimate. For brevity, the discussion of conventional classifiers is restricted to linear discriminant (LD), (Hand 1997; McLachlan 1992), and the resampling-weighted k-nearest neighbor (k-NN) (Steele and Patterson 2000) rules and a spatial classifier useful for land cover

mapping (Steele submitted). However, the combination methods are quite general, and other classifiers can be used in place of these classifiers. The probability estimates obtained from the jth classifier, $j = 1, \ldots, c$, are denoted by $p_h^j(x_0)$, $h = 1, \ldots, g$. The index j will be omitted when it is clear which classifier is being discussed.

3.1 The Resampling-Weighted *k*-NN Classifier

The resampling-weighted k-NN classifier is a smoothed version of the usual k-NN estimator (McLachlan 1992, Ch. 9) of $P[y_0 = h \mid t_0]$. The k-NN estimator of $P[y_0 = h \mid t_0]$ is the sample proportion of observations belonging to G_h among the k nearest neighbors and can be expressed as $\sum_{j=1}^{k} \Psi(\boldsymbol{y}_{0,j} = h)/k$, where $\Psi(P)$ is the indicator function of the event P and $\boldsymbol{y}_{0,j}$ denotes the group label of the jth closest observation to x_0 among $\boldsymbol{x}$. The weights are derived by taking the expectation of the k-NN estimator under bootstrap sampling. That is, the sample $\boldsymbol{x}$ is replaced by a random bootstrap sample $\boldsymbol{x}^*$ and the expectation is taken with respect to the empirical distribution function $\widehat{F}$. Accordingly, the resampling-weighted k-NN estimator is

$$\begin{aligned} p_h(x_0) &= \frac{1}{k}\sum_{j=1}^{k} E_{\widehat{F}}\Psi(\boldsymbol{y}_{0,j}^* = h) \\ &= \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{n} P_{\widehat{F}}(\boldsymbol{x}_{0,j}^* = \boldsymbol{x}_{0,i})\Psi(\boldsymbol{y}_{0,i} = h), \end{aligned} \tag{1}$$

where $\boldsymbol{y}_{0,j}^*$ denotes the group label of the jth closest observation to x_0 among $\boldsymbol{x}^*$ and $P_{\widehat{F}}(\boldsymbol{x}_{0,j}^* = \boldsymbol{x}_{0,i})$ is the probability that $\boldsymbol{x}_{0,i}$, the ith closest observation to x_0 among $\boldsymbol{x}$, is the jth closest sample observation to x_0 among $\boldsymbol{x}^*$. Rearranging equation (1) shows that the resampling estimator is a weighted average over all sample observations where the weight $w_{i,k}$ assigned to $\boldsymbol{x}_{0,i}$ is the probability that $\boldsymbol{x}_{0,i}$ will be among the k-closest neighbors of x_0. That is,

$$p_h(x_0) = \frac{1}{n}\sum_{j=1}^{n} w_{i,k}\, \Psi(\boldsymbol{y}_{0,i} = h),$$

where $w_{i,k} = \sum_{j=1}^{k} P_{\widehat{F}}(\boldsymbol{x}^*_{0,j} = \boldsymbol{x}_{0,i})/k$. In comparison, the conventional k-NN estimator assigns the weight $\Psi(i \leq k)\,/\,k$ to $\boldsymbol{x}_{0,i}$. Steele and Patterson (2000) present formulae for computing the $P_{\widehat{F}}(\boldsymbol{x}^*_{0,j} = \boldsymbol{x}_{0,i})$'s. For my examples, the accuracy estimates for the resampling weighted k-NN classifier were slightly larger than the conventional k-NN classifier. Generally, the resampling weighted k-NN classifier slightly faster to compute than the conventional k-NN classifier because tie-breaking is not necessary.

3.2 The Linear Discriminant Classifier

The linear discriminant classifier (Seber 1984, Ch. 4) is applied under the assumption of equal prior probabilities of group membership. The Mahalanobis distance from x_0 to G_h is denoted by $m_h(t_0) = \left(t_0 - \bar{t}_h\right)' \widehat{\boldsymbol{\Sigma}}^{-1} \left(t_0 - \bar{t}_h\right)$, where $\bar{t}_h$ denotes the (multivariate) sample mean of the group h covariate vectors, $\widehat{\boldsymbol{\Sigma}}$ is the pooled sample variance-covariance matrix, and u' denotes the transpose of the vector u. The posterior probability of membership in G_h is estimated by

$$p_h(x_0) = \frac{\exp[-m_h(t_0)/2]}{\sum_{j=1}^{g} \exp\left[-m_j(t_0)/2\right]}\,.$$

3.3 The Spatial Classifier

Variation in landform or climate may induce patterns in the spatial distribution of land cover types. For example, there will be differences in precipitation and vegetation distribution on opposing sides of a mountain range because of orographic effects. To exploit spatial information, Steele (submitted) proposed a classifier based on the spatial distances from an unclassified observation to its the spatially nearest training observations within each of the g groups. The spatial classifier is constructed as follows. The spatial point-to-group distance

from an observation x_0 to G_h, denoted by $d_h(z_0)$, is the Eculidean distance between z_0 and the location coordinate pair of its nearest training set neighbor in G_h. Similarly, $d_h(z_k)$ denotes the distance between z_k, the location of the kth training observation $x_k \in G_h$, and its nearest neighbor in G_h, and $\boldsymbol{D}_h = \{\, d_h(z_k) \mid k = 1, \ldots, n_h\}$ denotes the set of within-G_h nearest neighbor distances. Suppose that Z_0 denotes the location of a random map point. The rank of $d_h(z_0)$ among the elements of $\boldsymbol{D}_h$ is used to estimate the probability of obtaining a nearest neighbor distance greater than $d_h(z_0)$, given that the land cover type at Z_0 is G_h. Specifically, the estimator is

$$\widehat{P}[d_h(Z_0) > d_h(z_0) \mid y_0 = h] \tag{2}$$
$$= \begin{cases} \dfrac{|\{d_h(z_k) \in \boldsymbol{D}_h \mid d_h(z_k) > d_h(z_0)\}| - .5}{n_h}, & \text{if } d_h(z_0) < \max \boldsymbol{D}_h, \\ .5/\max\{n_1, \ldots, n_g\}, & \text{if } d_h(z_0) \geq \max \boldsymbol{D}_h. \end{cases}$$

where $|A|$ denotes the cardinality of the set A. The estimator of $P[d_h(Z_0) > d_h(z_0) \mid y_0 = h]$ reverses and scales the rank of $d_h(z_0)$ among $\boldsymbol{D}_h$ so that the small distances produce estimates near 1 and large distances produce estimates near 0. Finally, the spatial estimator of the unconditional probability of membership in G_h is defined to be

$$p_h(x_0) = \frac{\widehat{P}[d_h(Z_0) > d_h(z_0) \mid y_0 = h]}{\sum_{j=1}^{g} \widehat{P}[d_j(Z_0) > d_j(z_0) \mid y_0 = j]}. \tag{3}$$

Appendix 1 provides some additional details regarding the spatial classifier.

4. COMBINING CLASSIFICATION RULES

The problem of combining predictions originating from different predictors has been studied since the 1970's. Breiman (1996) and Hand (1997, Ch. 9) review the subject; recent articles include LeBlanc and Tibshirani (1996), Merz (1999), Merz and Pazzani (1999), and Mojirsheibani (1999). Rather than attempt to provide a comprehensive review of combination classifiers, I concentrate on one type of combination classifier and propose a simple alternative method. Suppose that c classifiers have produced estimates of $P[y_0 = h \mid x_0]$ for each

$h = 1, \ldots, g$. Arguably, the simpliest method of combining rules assigns x_0 to G_h if the sum $s_h(x_0) = \sum_{j=1}^{c} p_h^j(x_0)$ is largest among $s_1(x_0), \ldots, s_g(x_0)$. A weakness of this rule is that all values $p_h^1(x_0), \ldots, p_h^c(x_0)$ are given the same weight in the sum, yet there may be significant differences among classifiers with respect to estimation accuracy.

4.1 The Product Rule

An alternative method of combining classifiers, the *product rule*, is proposed which uses products

$$r_h(x_0) = \prod_{j=1}^{c} p_h^j(x_0), h = 1, \ldots, g, \tag{4}$$

instead of sums. The product rule estimator of $P[y_0 = h \mid x_0]$ is $r_h(x_0) / \sum_{k=1}^{g} r_k(x_0)$. An unclassified observation x_0 is assigned to G_h if $r_h(x_0)$ is largest among $r_1(x_0), \ldots, r_g(x_0)$. This rule imposes a form of consensus agreement among the classifiers in the sense that if just one classifier indicates that membership in G_h is very unlikely, then the combination classifier is unlikely to assign x_0 to G_h. The consensus property is a consequence of the product being smaller than its smallest component term. Specifically, if $p_h^k(x_0)$ is near zero, then $r_h(x_0)$ is also near zero because $r_h(x_0) \leq p_h^k(x_0)$, $k = 1, \ldots, c$. The product rule is very easy to compute once the conventional classifiers have produced the estimates of $P[y_0 = h \mid x_0]$. In contrast, other combination methods, including the stacked regression method discussed below, require substantially more computational effort.

4.2 Stacked Regression

The stacked regression method (Breiman 1996; Wolpert 1992) uses the method of least squares to try to find linear combinations of the probability estimates from several classifiers which improve on individual classifier accuracy. To set up the discussion, suppose that membership probabilities are estimated for x_0 using c different classifiers, and all gc estimates are collected in the vector

$$\underset{gc\times 1}{p(x_0)} = [\, p_1^1(x_0) \cdots p_g^1(x_0) \cdots p_1^c(x_0) \cdots p_g^c(x_0)\,]'. \tag{5}$$

The vectors $p(x_i)$, $i = 1, \ldots, n$, are stacked as rows in a $n \times gc$ matrix denoted by $\boldsymbol{P}$. Let u_h denote a n-vector identifying training set memberships in G_h; i.e., the ith element is

$$u_{i,h} = \begin{cases} 1, & \text{if } y_i = h, \\ 0, & \text{if } y_i \neq h. \end{cases} \tag{6}$$

Consider the model $u_h = \boldsymbol{P}\beta_h + \varepsilon_h$, where β_h is a gc-vector of unknown parameters, and ε_h is a vector of independent random variables with mean 0 and a common variance. Then, an estimate $\widehat{\beta}_h$ can be obtained by solving the normal equations $\boldsymbol{P}'\boldsymbol{P}\beta_h = \boldsymbol{P}'u_h$ for β_h. An unclassified observation x_0 is assigned to the group with the largest value of $\widehat{u}_{0,h} = p(x_0)'\widehat{\beta}_h$, $h = 1, \ldots, g$. Tthe columns of $\boldsymbol{P}$ are subject to c linear constraints because the estimated group membership probabilities obtained from a particular classifier are constrained to sum to 1. Hence, $\boldsymbol{P}'\boldsymbol{P}$ is singular. A solution to the normal equations is $\widehat{\beta}_h = (\boldsymbol{P}'\boldsymbol{P})^-\boldsymbol{P}'u_h$, where $(\boldsymbol{P}'\boldsymbol{P})^-$ is the Moore-Penrose generalized inverse (Schott 1997, Ch. 5) of $\boldsymbol{P}'\boldsymbol{P}$.

Over-fitting is a potentially important problem with the stacked regression approach as described above. The β_h's will tend to be over-fit (Breiman 1996; LeBlanc and Tibshirani 1996) because the u_h's are regressed on the columns of $\boldsymbol{P}$, and $\boldsymbol{P}$ is determined by the u_h's. Breiman (1996) and LeBlanc and Tibshirani (1996) reduce over-fitting by constructing $\boldsymbol{P}$ from cross-validation versions of the $p(x_i)$'s. Specifically, x_i is held-out in calculating $p(x_i)$. The remaining $n - 1$ observations and all c classifiers are used to construct c new classification rules and to compute the estimates $p_1^1(x_i), \ldots, p_g^c(x_i)$, $i = 1, \ldots, n$. For large data sets, computing $\boldsymbol{P}$ by cross-validation is time-consuming because nc classification rules must be constructed.

An alternative stacked regression model, motivated by the product rule, is $u_h = \ln \boldsymbol{P}_h\gamma_h + \varepsilon_h^*$ where the $n \times c$ matrix $\boldsymbol{P}_h$ is constructed from the c columns of $\boldsymbol{P}$ containing the probability estimates of group membership in G_h. The estimates $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_g$ can be used to construct a product rule which computes $\prod^c p_h^j(x_0)^{\widehat{\gamma}_j}$ instead of formula (4). This strategy was

found to perform poorly relative to both the product rule and the stacked regression methods and will not be discussed further.

4.3 Improvements to Stacked Regression

A second problem, not unique to stacked regression, is that least squares may not yield the best possible predictor of the membership label y_0. LeBlanc and Tibshirani (1996) and Breiman (1996) discuss several approaches to improving predictions. Three approaches were investigated. The first, subset regression, uses a subset of the predictors (i.e., columns of $\boldsymbol{P}$). Subset regression was carried out using a backwards selection algorithm which retained a predictor only if the ratio of corresponding parameter estimate to its standard error estimate was at least 2 in absolute value. A second approach imposes non-negativity constraints on the components of $\widehat{\beta}_h$ to insure that the $\widehat{u}_{0,h}$'s are non-negative. Following LeBlanc and Tibshirani (1996) and Breiman (1996), Lawson and Hanson's (1974) algorithm was used to impose non-negativity constraints.

The third approach to bias correction is LeBlanc and Tibshirani's (1996) method of computing $\widehat{\beta}$ from bias-corrected versions of $\boldsymbol{P}'\boldsymbol{P}$ and $\boldsymbol{P}'u_h$. Their method draws B bootstrap training samples, $\boldsymbol{x}^{*b}$, $b = 1, \ldots, B$ from $\boldsymbol{x}$. The bootstrap samples are used to calculate group membership probabilities for each $x_i \in \boldsymbol{x}$ and $x_j \in \boldsymbol{x}^{*b}$ using classifiers constructed from $\boldsymbol{x}^{*b}$. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be subsets of $\boldsymbol{x}$, and let $\boldsymbol{P}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ denote a matrix of probability estimates computed by using $\boldsymbol{x}_1$ to classify $\boldsymbol{x}_2$. Let $u_h(\boldsymbol{x}_2)$ denote the indicator vector identifying membership in G_h for the observations contained in $\boldsymbol{x}_2$. The bootstrap corrected versions of $\boldsymbol{P}'\boldsymbol{P}$ and $\boldsymbol{P}'u_h$ are

$$(\boldsymbol{P}'\boldsymbol{P})_{bc} = \boldsymbol{P}(\boldsymbol{x}, \boldsymbol{x})'\boldsymbol{P}(\boldsymbol{x}, \boldsymbol{x}) + \frac{1}{B}\sum_{b=1}^{B}[\boldsymbol{P}(\boldsymbol{x}^{*b}, \boldsymbol{x})'\boldsymbol{P}(\boldsymbol{x}^{*b}, \boldsymbol{x}) - \boldsymbol{P}(\boldsymbol{x}^{*b}, \boldsymbol{x}^{*b})'\boldsymbol{P}(\boldsymbol{x}^{*b}, \boldsymbol{x}^{*b})]$$

and

$$(\boldsymbol{P}'u_h)_{bc} = \boldsymbol{P}(\boldsymbol{x}, \boldsymbol{x})'u_h(\boldsymbol{x}) + \frac{1}{B}\sum_{b=1}^{B}[\boldsymbol{P}(\boldsymbol{x}^{*b}, \boldsymbol{x})'u_h(\boldsymbol{x}) - \boldsymbol{P}(\boldsymbol{x}^{*b}, \boldsymbol{x}^{*b})'u_h(\boldsymbol{x}^{*b})].$$

Bias-corrected estimators of β_h are corrected by solving the normal equations $(\boldsymbol{P}'\boldsymbol{P})_{bc}\beta_h = (\boldsymbol{P}'u_h)_{bc}$. A particular advantage of this approach is that variable selection is easily accomplished by deleting rows and columns of $(\boldsymbol{P}'\boldsymbol{P})_{bc}$ and $(\boldsymbol{P}'u_h)_{bc}$ before solving the normal equations.

I propose a bias avoidance approach based on the leave-one-out bootstrap (Efron and Tibshirani 1997). To set up the method, let $\boldsymbol{x}_{-}^{*b}$ denote set of observations that have been left out of the boostrap sample (approximately 36.8% are left out of $\boldsymbol{x}^{*b}$). Let $\boldsymbol{P}(\boldsymbol{x}^{*b},\boldsymbol{x}_{-}^{*b})$ denote the matrix obtained from constructing the classifiers from $\boldsymbol{x}^{*b}$, and estimating group membership probabilities for each observation in $\boldsymbol{x}_{-}^{*b}$. The leave-one-out versions of $\boldsymbol{P}'\boldsymbol{P}$ and $\boldsymbol{P}'u_h$ are

$$(\boldsymbol{P}'\boldsymbol{P})_{-} = \frac{1}{B}\sum_{b=1}^{B}[\boldsymbol{P}(\boldsymbol{x}^{*b},\boldsymbol{x}_{-}^{*b})'\boldsymbol{P}(\boldsymbol{x}^{*b},\boldsymbol{x}_{-}^{*b})$$

and

$$(\boldsymbol{P}'u_h)_{-} = \frac{1}{B}\sum_{b=1}^{B}[\boldsymbol{P}(\boldsymbol{x}^{*b},\boldsymbol{x}_{-}^{*b})'u_h(\boldsymbol{x}_{-}^{*b}).$$

Variable selection can be carried out using the leave-one-out version of $\boldsymbol{P}'\boldsymbol{P}$ and $\boldsymbol{P}'u_h$. Computationally, the bias avoidance approach is somewhat easier to implement and faster to carry out than LeBlanc and Tibshirani's (1996) bias correction approach because it is not necessary to compute $\boldsymbol{P}(\boldsymbol{x},\boldsymbol{x})$.

4.4 Mojirsheibani's (1999) Method

Mojirsheibani (1999) combines multiple classification rules by identifying training observations that are the same as x_0 with respect to predicted group membership. This subset of training observations are used classify x_0 by finding the plurality group. Specifically, each training observation, x_i, $i = 1,\ldots,n$, is classified using each of the c classifiers and the training set of $n-1$ observations obtained by holding out x_i. Let $\widehat{\boldsymbol{y}}_i = (\widehat{y}_i^1,\ldots,\widehat{y}_i^c)$ and $\widehat{\boldsymbol{y}}_0$ denote the c-tuples of membership predictions for x_i and for x_0, respectively, obtained from the c

classifiers. Mojirsheibani's (1999) method forms the subset of training observations which are classified the same as x_0 by all classifiers, i.e., $\boldsymbol{s}_0 = \{ x_i \in \boldsymbol{x} \mid \widehat{y}_i^1 = \widehat{y}_0^1, \ldots, \widehat{y}_i^c = \widehat{y}_0^c \}$, and predicts x_0 to belong to the group with greatest number of training observations in $\boldsymbol{s}_0$.

5. Accuracy Estimation

Accuracy was estimated using k-fold cross-validation (Efron and Tibshirani 1993, Ch. 17). A k-fold cross-validation splits the data into k disjoint subsets of approximately equal size. The bth subset, $b = 1, \ldots, k$, is classified using the remaining $k - 1$ subsets, and the predictions are compared to the recorded memberships to obtain accuracy estimates. For example, n-fold cross-validation classifies each observation using the rule constructed from the other $n - 1$ observations. For stacked regression methods, an exact n-fold cross-validation algorithm requires that all $n - 1$ vectors $p(x_j)$, $j = 1, \ldots, i-1, i+1, \ldots, n$ be recomputed when x_i, $i = 1, \ldots, n$, is left out; consequently, $cn(n-1)$ classification rules must be computed. If all $n - 1$ vectors $p(x_j)$ are not recomputed, x_i will contribute to each of the $p(x_j)$'s, albeit weakly. Hence, x_i will influence $\boldsymbol{P}$ and the estimates of $\beta_1, \ldots, \beta_g$, through the $p(x_j)$'s. Ten-fold cross-validation was used in this study because the enormous computational effort demanded by n-fold cross-validation for stacked regression and the example data sets. An approximate cross-validation algorithm which did not recompute $p(x_j)$ when classifying x_i yielded accuracy estimates for stacked regression combination classifiers that were sometimes substantially larger than the 10-fold cross-validation estimates. For the example data sets discussed herein, and probably in general, it is necessary to use disjoint training and test sets when estimating accuracy for combination classifiers because of the tendency for over-fitting.

6. Examples

In this section, we discuss classification results for two data sets used to classify adjacent Landsat TM scenes Path 41, Row 29 (P41/R29) and P41/R28. These scenes cover a combined land area of 6.42 million hectares located in west-central Montana and central Idaho and encompass a rectangular region approximately 75 by 300 km in dimension (Figure 1). The land cover type maps were prepared jointly by the Craighead Wildlife-Wildlands Institute,

Missoula MT, and the Wildlife Spatial Analysis Laboratory, Montana Cooperative Wildlife Research Unit, of the University of Montana for the purpose of analyzing habitat suitability for a number of species such as grizzly bears and bighorn sheep. Craighead et al. (1999) provide a detailed discussion of the research.

The landscape of the northern scene (P41/R28) is dominanted by the Bitterroot mountains, and the climate is moist Pacific maritime moderated by occasional intrusions of continental air masses (Cooper et al., 1987). The vegetation is generally mesic; montane forests dominate the region but there are lesser amounts of shrublands and grasslands. Agricultural and urban areas occupy only a few percent of the region. The southern scene (P41/R29) is dominanted by the Idaho batholith, a large, uplifted mountainous region located in the north and center, and basin and range landscapes in the south. The climate changes rapidly from Pacific maritime in the extreme north to arid Basin and Range in the south (Steele et al., 1981). Vegetation is a mosaic of forest, shrub, and grassland community types with agriculture largely limited to the Snake River plain located along the southern periphery of the scene. Table 1 shows a list of land cover types used for mapping purposes.

The combined scenes contain more than 4.8 million hectares of wilderness and adjacent unroaded areas, almost all of which is under Federal ownership. Because of the cost and effort required of sampling in mountain wildernesses, the training sets were constructed by combining data originating from a variety of sampling projects. Usually, these projects were conducted by government agencies for land cover mapping purposes, but some of the data were collected to simply summarize the range and extent of vegetation types. USDA Forest Service personnel usually sampled accessible areas within their Forest or District boundaries, but the Craighead Wildlife-Wildlands Institute concentrated their sampling efforts in and around the remote and relatively inaccessible wilderness areas. Privately held lands were rarely sampled. Figure 1, showing plot locations and topographic relief for the combined scenes reveals that sampling intensity and plot (sample point) density is spatially variable with several regions with few or no plots and a few regions with high plot densities. Obviously, these data do not constitute a

probability sample; hence, accuracy estimates obtained from the training data are only meaningful for the sampled areas.

The land cover type classification system for Landsat TM scene P41/R28 identified 15 land cover types (described in Table 1), and 3005 training observations were used for classification. The covariates used for classification were seven of the eight TM bands, elevation, slope, and a measure of solar insolation computed from aspect and slope. Three classifiers were used: the 20-resampling weighted k-NN rule, the linear discriminant (LD) rule, and the spatial classifier, and combined pair-wise and as a triple.

6.1 Combination Classifiers

Ten-fold cross-validation accuracy estimates were computed using the product rule, Mojirsheibani's (1999) method, and several variants of stacked regression. The variations of stacked regression that with and without variable selection and no bias correction, stacked regression with non-negativity constraints and without variable selection and no bias correction, stacked regression with variable selection and with LeBlanc and Tibshiriani's bias correction, and stacked regression with variable selection and the bias avoidance method. Four combinations were produced using each of the combination methods: 20-NN resampling & LD, 20-NN resampling & spatial, LD & spatial, 20-NN resampling & LD & spatial. As a matter of organization, those methods that do not involve bootstrap algorithms are compared first and separately from those methods utilizing the bootstrap.

Overall cross-validation accuracy estimates for 20-NN, LDF, and spatial classifier were 0.569, 0.552, and 0.446, respectively. Table 2 compares 10-fold cross-validation accuracy estimates obtained from TM scene P41/R29.

6.1 Combination Classifiers

Overall cross-validation accuracy estimates for 20-NN, LDF, and spatial classifier were 0.569, 0.552, and 0.446, respectively. Table 2 shows the overall CV accuracy estimates for three combination classifiers: 20-NN and the spatial classifier, LDF and the spatial classifier, and 20-NN, LDF and the spatial classifier, and six combination methods: the class majority,

sum and product rules, and stacked regression with and without non-negativity constraints, and stacked log-regression. The class majority, sum, and product rules, and constrained stacked regression yielded similar overall CV estimates ranging from 0.597 to 0668. The unconstrained stacked regression and stacked log-regression methods yielded larger CV accuracy estimates, ranging from 0.682 to 0.826. The best combined CV accuracy estimates corresponds to an improvement of $45\% = 100(0.826 - 0.569)/0.569$ over the best single classifier CV accuracy estimate. Table 3 shows CV accuracy estimates for each group obtained from the uncombined classifers and the three combination classifiers using stacked log-regression to combine the classifiers. In a few instances, single classifiers produced larger CV estimates than the combined classifers (e.g., the CV estimates for 3310 using LDF and the spatial classifier is 0.00. Table 4 compares group CV accuracy estimates obtained from the combined 20-NN, LDF and the spatial classifier by combination method.

6. Conclusions

Acknowledgments

Funding was provided by the USDA Forest Service, Boise and Custer National Forests, the USGS Biological Resources Division, Gap Analysis Program, Montana Department of Fish, Wildlife, and Parks, and The University of Montana. We could not have undertaken this research without extensive help and cooperation from our colleagues Chris Winne and Chip Fisher at the Wildlife Spatial Analysis Lab. Finally, we thank the numerous individuals from the Beaverhead, Custer, and Salmon/Challis National Forests who not only collected ground reference data but also made them available to us for this research.

References

ANDERSON, J.R., HARDY, E.E., ROACH, J.T., WITMER, R.E., 1976, A land use and land cover classification for use with remote sensor data. *U.S. Geological Survey Professional Paper* 964. Washington, D.C.: U.S. Govt Printing Office.

CANTERS, F., 1997, Evaluating the uncertainty of area estimates derived from fuzzy land-cover classification. *Photogrammetric Engineering and Remote Sensing*, **63**, 403-414.

CONGALTON, R.G., 1988, Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, **54**, 587-592.

CONGALTON, R.G., 1991, A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing and Environment*, **37**, 35-46.

Cooper, S. V., K. E. Neiman, D. W. Roberts. 1987. Forest habitat types of northern Idaho: A second approximation. General Technical Report INT-236. U.S. Department of Agriculture, Forest Service, Intermountain Research Station, Ogden, UT. 143 pp.

CRESSIE, N.A.C., 1993, *Statistics for Spatial Data* (New York: Wiley).

DUDANI, S.A., 1976, The distance-weighted k-nearest-neighbor rule. *I.E.E.E. Transactions on Systems, Man and Cybernetics*, **6**, 325-327.

EFRON, B., 1983, Estimating the error rate of a prediction rule: Improvement on cross-validation *Journal of the American Statistical Association*, **78**, 316-331.

EFRON, B. and TIBSHIRANI, R., 1993, *An Introduction to the Bootstrap* (London: Chapman and Hall).

FISHER, F.B., WINNE, J.C., THORTON, M.M., TADY, T.P., MA, Z., HART, M.M., and REDMOND, R.L., 1998, Atlas of Montana Land Cover. Unpublished Appendix to the Montana Gap Analysis Final Report; submitted to the USGS, Biological Resources Division, National Gap Analysis Program, Moscow, ID.

GEMAN, S. and GEMAN, D., 1984, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *I.E.E.E. Transactions on Pattern Recognition and Machine Intelligence*, **6**, 721-741.

Foody, G.M., McCulloch, M.B., Yates, W.B. (1995), "Classification of Remotely Sensed Data by an Artifical Neural Network: Issues Related to Training Data Characteristics," *Photogrammetric Engineering and Remote Sensing*, **61**, 391-401.

HASLETT, J., 1985, Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context. *Pattern Recognition*, **18**, 287-296.

HJORT, N.L. and MOHN, E., 1984, A comparison of some contextual methods in remote sensing. *Proceedings of the Eighteenth International Symposium on Remote Sensing of the Environment*, Centre National d'Etudes Spatiales, Paris, France, pp. 1693-1702.

Homer, C.G., Ramsey, R.D., Edwards, T.C., and Falconer, A. (1997), "Landscape Cover-Type Modeling Using a Multi-Scene Thematic Mapper Mosaic," *Photogrammetric Engineering and Remote Sensing*, **63**, 59-67.

HUBERTY, C.J., 1994, *Applied Discriminant Analysis* (New York: Wiley).

KARTIKEYAN, B., GOPALAKRISHNA, B., KALABURME, M.H., and MAJUMDAR, K.L., 1994, Contextual techniques for classification of high and low resolution remote sensing data. *International Journal of Remote Sensing* **15**, 1037-1051.

LACHENBRUCH, P.A. and MICKEY, M.R., 1968, Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1-11.

LeBlanc, M. and Tibshiriani, R. (1996), "Combining Estimates in Regression and Classification," *Journal of the American Statistical Association*, **91**, 1641-1650.

LEHMAN, E.L., 1975, *Nonparametrics: Statistical Methods Based on Ranks* (San Francisco: Holden-Day).

McLACHLAN, G.J., 1992, *Discriminant Analysis and Statistical Pattern Recognition* (New York: Wiley).

MACLEOD, J.E.S., LUK, A., and TITTERINGTON, D.M., 1987, A re-examination of the distance-weighted k-nearest-neighbor classification rule. *I.E.E.E. Transactions on Systems, Man and Cybernetics*, **17**, 689-696.

PRESS, S.J., 1996, The directional neighborhoods approach to contextual classification of images from noisy data. *Journal of the American Statistical Association*, **91**, 1091-1100.

Scott, J.M., Davis, F., Csuti, B., Noss, R., Butterfield, B., Caicco, S., Groves, C., Edwards, T.C., Jr., Ulliman, J., Anderson, H., Derchia, F., and Wright, R.G. (1993) "Gap Analysis: A Geographic Approach to Protection of Biological Diversity," *Wildlife Monographs*, 123.

SEBER, G.A.F., 1984, *Multivariate Observations* (New York: Wiley).

SHARMA, K.M.S. and SARKAR, A., 1998, A modified contextual classification technique for remote sensing data. *Photogrammetric Engineering and Remote Sensing*, **64**, 273-280.

SILVERMAN, B.W., 1986, *Density Estimation for Statistics and Data Analysis* (London: Chapman and Hall).

STEELE, B.M. and PATTERSON, D.A., (1998), Ideal bootstrap estimation of expected prediction error for k-nearest neighbor classifiers: applications for classification and error assessment. Submitted to *Statistics and Computing*.

STEELE, B.M., WINNE, J.C., and REDMOND, R.L., (1998), Estimation and mapping of misclassification probabilities for thematic land cover maps. To appear in Remote Sensing and Environment.

Steele, R., R. D. Pfister, R. A. Ryker, J. A. Kittams. 1981. Forest habitat types of central Idaho. General Technical Report INT-114. U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station, Ogden, UT. 138 pp.

STEHMAN, S.V. and CZAPLEWSKI, R.L., 1998, Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing and Environment*, **64**, 331-334.

VAN DEUSEN, P.C., 1995, Modified highest confidence first classification. *Photogrammetric Engineering and Remote Sensing*, **61**, 419-425.

WESTFALL, P.H. and YOUNG, S., 1992, *Resampling-Based Multiple Testing* (New York: Wiley).

APPENDIX

When $d_h(z_0) \geq \max \boldsymbol{D}_h$, the estimate of $P(d_h(Z_0) > d_h(z_0) \mid y_0 = h)$ is defined to be $.5/\max\{n_1, \ldots, n_g\}$ for the following reasons. Suppose that x_0 is distant from all training points; specifically, suppose that $d_h(z_0) \geq \max \boldsymbol{D}_h$ for $h = 1, \ldots, g$. Then, there is little or no spatial information indicating that membership in one group is more likely than another, and $\widehat{P}(d_h(Z_0) > d_h(z_0) \mid y_0 = h)$ should take on the same value for each $h = 1, \ldots, g$. The value $.5/\max\{n_1, \ldots, n_g\}$ is used so that $\widehat{P}(d_h(Z_0) > d_h(z_0) \mid y_0 = h)$ is the same for each h and to insure that it is a monotonically decreasing function of $d_h(z_0)$. This definition also accounts for sample size differences between groups. Point-to-nearest neighbor distances will tend to be large for groups with small sample sizes, and vice versa. Consequently, without adjustments for sample size differences, the spatial classifier will tend to assign x_0 to the groups with the largest sample sizes. Definition (0) may be modified to incorporate prior probabilities via the usual Bayes rule calculation of posterior probabilities (McLachlan 1992, Ch. 2). Moreover, the spatial classifier may be modified by defining the point-to-distance d_h differently; e.g., $d_h(z_0)$ could be defined as the mean distance from z_0 to its k nearest neighbors in G_h.

Tables

Table 1. Land cover types, sample sizes, and LD 10-fold cross-validation accuracy estimates for LANDSAT scenes P41/R28 and P41/R29.

		P41/R28		P41/R29	
Code	Land cover type	N	Accuracy	N	Accuracy
3100	Xeric grassland	89		409	0.83
3180	Mesic grassland and subapline meadows	57		31	0.35
3200	Mesic Shrubland	272		316	0.58
3300	Xeric Shrubland	–		388	0.70
3310	Non-sagebrush xeric shrubland	38		–	
3350	Sagebrush shrubland	30		–	
3400	Shrub and herbacious-dominated burn	49		111	0.62
4203	Lodgepole pine dominated forest	472		631	0.70
4206	Ponderosa pine dominated forest	100		153	0.55
4212	Douglas fir dominated forest	431		502	0.49
4230	Mixed Douglas fir/ponderosa pine	–		322	0.42
4260	Mixed whitebark pine forest	135		221	0.72
4270	Mixed subapline forest	497		243	0.55
4280	Mesic forest conifer regeneration dominated burn	394		183	0.63
4410		69		59	0.61
6200	Herbaceous riparian	33		80	1.00
7300	Rock dominated	112		225	0.67
9100	Snowfield or ice	–		19	0.79
	Total	2778		3923	0.6345