Can. J. Remote Sensing, Vol. 39, No. 6, pp. 507–520, 2013

A new Bayesian ensemble of trees approach for land cover classification of satellite imagery

Reshu Agarwal, Pritam Ranjan, and Hugh Chipman

Abstract. Classification of satellite images is a key component of many remote sensing applications. One of the most important products of a raw satellite image is the classification that labels image pixels into meaningful classes. Though several parametric and nonparametric classifiers have been developed thus far, accurate classification still remains a challenge. In this paper, we propose a new reliable multiclass classifier for identifying class labels of a satellite image in remote sensing applications. The proposed multiclass classifier is a generalization of a binary classifier based on the flexible ensemble of regression trees model called Bayesian Additive Regression Trees. We used three small areas from the LANDSAT 5 TM image, acquired on 15 August 2009 (path–row: 08–29, L1T product, UTM map projection) over Kings County, Nova Scotia, Canada, to classify the land cover. Several prediction accuracy and uncertainty measures have been used to compare the reliability of the proposed classifier with the state-of-the-art classifiers in remote sensing.

Résumé. La classification des images satellitaires est une composante essentielle d'un grand nombre d'applications en télédétection. Un des produits les plus importants d'une image satellitaire brute est sa classification qui sépare les pixels de l'image en classes utilisables. Bien que plusieurs méthodes paramétriques et non paramétriques aient été développées à ce jour, la précision de la classification demeure un défi. Dans cet article, nous proposons une nouvelle méthode de classification multiclasse pour l'identification des classes présentes dans une image satellitaire pour des applications en télédétection. L'approche est une généralisation des classificateurs binaires fondée sur le modèle d'ensemble flexible d'arbres de régression bayésienne. Nous avons utilisé trois petites zones de l'image LANDSAT 5 TM, acquise le 15 aout 2009 (« path–row »: 08–29, produit L1T, projection UTM) sur Kings County, Nova Scotia, Canada, pour la classification de l'occupation du sol. Plusieurs mesures de la précision des prévisions et de l'incertitude ont été utilisées pour comparer la fiabilité du classificateur proposé avec les classificateurs courants en télédétection.
[Traduit par la Rédaction]

Introduction

Satellite image classification plays a crucial role in numerous remote sensing data applications such as land use planning, land cover change monitoring, forest degradation assessment, hydrological modeling, sustainable development, wildlife habitat modeling, biodiversity conservation, and so on. One of the most important inputs in such an application is classification, in which each pixel receives a class label. Thus, it becomes essential to achieve the highest possible accuracy of the classified maps. Several classifiers have been developed and implemented worldwide (e.g., Wacker and Landgrebe, 1972; Anderson et al., 1976; Jensen, 1996; Franklin et al., 2002; Pal and Mather, 2003; Gallego, 2004); however, accurate labeling of the pixels still remains a challenge often due to shadows, spectral mixing and overlapping classes.

Among various parametric classification methods, the maximum likelihood (ML) classifier has been the most widely used classifier in remote sensing image processing software (Peddle, 1993). However, the ML classifier is based on a parametric model that assumes normally distributed data which is often violated in complex land cover (or land use) satellite images (Lu and Weng, 2007). Nonparametric classifiers, which do not make strong assumptions such as normality, have gained much popularity. Classifiers based on k-nearest neighbor (k-NN), artificial neural networks (ANN), decision trees, and support vector machines (SVM) have shown better performance compared with ML classifiers for complex landscapes (Zhang and Wang, 2003; Bazi and Melgani, 2006; Li et al., 2010). Of course, the nonparametric methods are not perfect either and have many shortcomings. Ranking of such classifiers has been an interesting research area in the machine learning literature. For example, Sudha and Bhavani (2012) concluded that SVM is a better classifier than k-NN, and Song et al. (2012) argued that SVM is either comparable with, or slightly better than, ANN.

Decision tree based classifiers became extremely popular in machine learning after classification and regression trees (CART) were introduced by Breiman et al. (1984), although this type of classifier had been around since the 1960s under

Received 28 October 2012. Accepted 19 December 2013. Published on the Web at http://pubs.casi.ca/journal/cjrs on 5 March 2014.

Reshu Agarwal, Pritam Ranjan[1], and Hugh Chipman. Department of Mathematics and Statistics, Acadia University, 15 University Ave., Wolfville, NS B4P 2R6, Canada.

[1]Corresponding author:(e-mail: pritam.ranjan@acadiau.ca).

the name of concept learning systems. In remote sensing applications, CART has been successfully used for the classification of multispectral and hyper-spectral images with high accuracy (e.g., Hansen et al., 1996; Friedl and Brodley, 1997; Yang et al., 2003). Refinements over CART (e.g., bagging, boosting, and random forests) have also been used in remote sensing for more accurate class label identification (e.g., Lawrence et al., 2004). An illustration in this paper indicates that CART can sometimes yield unreliable predictions.

In this paper, we propose a new reliable multiclass classifier for accurate class label prediction. The proposed classifier is based on the Bayesian "ensemble of trees" model called Bayesian Additive Regression Tree (BART) originally developed by Chipman et al. (2010). In the context of drug discovery and credit risk modeling, BART has been used for constructing binary classifiers under the name of BART probit (Chipman et al., 2010) and Bayesian Additive Classification Tree (BACT) (Zhang and Hardle, 2010). In this paper, we follow a one-against-all approach for generalizing this binary classifier to a multiclass classifier that is referred to as mBACT. As illustrated in the results section, mBACT yields more reliable predicted class labels than two popular competing classifiers, SVM and CART.

For performance comparison of different classifiers, we use a LANDSAT 5 TM image covering Kings County of Nova Scotia, Canada, acquired on 15 August 2009. This is a predominantly rural area with several small towns. The satellite (LANDSAT 5 TM) stopped working in November 2011 and the images over Kings County from September 2009 to November 2011 were unclear (i.e., cloudy or snowy). Thus, we used the scene acquired on 15 August 2009 (path–row: 08–29, L1T product, UTM map projection) for performance comparison. The satellite image consists of six reflectance bands (blue, green, red, near infrared, and two middle infrared) with 30 m resolution and one thermal band with 120 m resolution. For maintaining resolution consistency, the thermal band was not used in building the classifier. The LANDSAT scene is 185 km $\times$ 170 km, and our study area consists of three relatively small regions in Kings County, i.e., the towns of Wolfville, Windsor, and Kentville and their surrounding areas. Each image consists of six reflectance matrices (corresponding to the six bands), and each pixel can be classified in one of the seven land-use classes, built-up, pond–lake–river water, Bay of Fundy, agricultural land, grassland, trees, and scrubland.

The next section presents brief reviews on the methodologies and implementation details (in R software) of SVM, CART, BART, and BART probit models. We propose the new classification methodology, mBACT, in the following section. Note that the two methodology sections (Background and New Methodology) may appear to be somewhat terse and theoretical and can be skipped if only interested in the application and (or) performance comparison of the classifiers. The following section discusses the data set obtained from LANDSAT 5 TM image and study areas. A brief review of the accuracy and uncertainty measures used for performance comparison are then presented. Then the classified images and tabulated results for the proposed and competing classifiers are presented. Finally, the paper concludes with the overall comparison of all classifiers, and a few remarks.

Background

In this section we briefly review the key ideas of SVM and CART, and we present the necessary details of BART methodologies that are relevant for the development of mBACT. For details on SVM, CART, BART, and BACT see Liu and Zheng (2005), Breiman et al. (1984), Chipman et al. (2010), and Zhang and Hardle (2010), respectively.

Support vector machine (SVM)

SVM was originally developed by Vapnik (1995) for binary classification, and was later extended to the domain of regression problems (Vapnik et al., 1996). A basic SVM classifier (designed for binary response) takes a set of inputs and predicts the class label for every given input. The main idea is to construct a separating hyperplane in the input space that can divide the training data into two classes with minimal error.

Let Y be the response variable and $X = \{X^1, X^2, \ldots, X^p\}$ be the set of p independent predictor variables. Suppose D_0 consists of N training data points with binary response $y_i \in \{-1, 1\}$ at p-dimensional input $x_i = (x_i^1, x_i^2, \ldots, x_i^p)$, for $i = 1, \ldots, N$. Then, the linear separating hyperplane in the input space is $\{x: f(x) = w^T x + b = 0\}$, where $w \in R^p$, b is a scalar and $f(x)$ is the decision function. That is, $\hat{f}(x) > 0$ implies that the predicted class label at x is $\hat{y}(x) = 1$, whereas $\hat{f}(x) < 0$ supports $\hat{y}(x) = -1$. The parameters (w, b) are obtained by solving the following optimization problem:

$$\begin{aligned} &\text{Minimize } L(w) = \frac{1}{2}||w||^2 \\ &\text{Subject to } y_i f(x_i) \geq 1,\ i = 1, \ldots, N \end{aligned} \tag{1}$$

A linear hyperplane is often insufficient for separating the data into two classes, and a nonlinear hyperplane has to be constructed. This is achieved by transforming the input data in a much higher (possibly infinite) dimensional space called the "feature space" $((x \rightarrow \phi(x)))$, and finding a linear hyperplane in that space. Typically, a kernel function $K(x_i, x_j)$ is used to implicitly define this nonlinear transformation. Thus, the decision function becomes

$$f(x) = w^T \phi(x) + b = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b$$

where α_i and b are estimated by solving the optimization problem in Equation (1) with this decision function. A few popular kernels are as follows:

- Polynomial kernel: $K(x_i, x_j) = \left(\gamma\left\langle x_i, x_j\right\rangle + r\right)^d$
- Gaussian kernel: $K(x_i, x_j) = \exp\left(-\sigma|x_i - x_j|^2\right)$
- Sigmoid kernel: $K(x_i, x_j) = \tanh\left(\gamma\left\langle x_i, x_j\right\rangle + r\right)$

where γ and σ are scale parameters, r is an offset parameter, and d is the degree of the polynomial kernel.

In this paper, we are interested in a multiclass classification problem with n = seven classes (built-up, water, Bay of Fundy, agricultural land, grassland, trees and scrubland). The data set D consists of N points with response $y_i \in \{1, 2, \ldots, n\}$ for every p-dimensional input $x_i = (x_i^1, x_i^2, \ldots, x_i^p)$. One of the most popular technique for classifying multiclass data is to use the one-against-all approach (e.g., Bottou et al., 1994; Liu and Zheng 2005), where the main idea is to build a series of n independent decision functions, f_k (x), $k = 1, \ldots, n$ and then choose the class label that corresponds to the largest f_k, i.e., $\hat{y}(x) = \arg\max\{f_k(x), k = 1, \ldots, n\}$. Alternatively, one can follow the one-against-one approach (e.g., Knerr et al., 1990; Friedman, 1996), in which all pairwise, $n(n–1)/2$, binary classifiers are trained and the class with maximum votes is the predicted class label $\hat{y}(x)$. Hsu and Lin (2002) argue that the one-against-one approach can often outperform the one-against-all method.

We used the built-in function ksvm () in the R package "kernlab" (Karatzoglou et al., 2013) for fitting all SVM models. The classifier uses the one-against-one approach for solving a multiclass-classification problem. We used most of the default arguments in ksvm () (including type = "C-svc"), except the kernel parameters specified by kernel = "polydot" and kpar = list (degree = 2, scale = 1, offset = 0.5). These values of the parameters in "kpar" have been chosen based on a preliminary study of the datasets considered in this paper. One may find alternative parameters combination to be more appropriate in other applications. As expected from any classifier, ksvm () can produce both the predicted class label $\hat{y}(x)$ and the classification probability $\hat{p}_k(x) = \hat{p}(Y = k|x)$ for every class $k = 1, \ldots, n$. These values are achieved by using type = "probabilities" and type = "response" in the built-in function predict.svm().

It is worth noting that the predicted class label obtained via ksvm() does not necessarily match with the maximum predicted classification probability $\hat{p}_{\max}(x) = \max\{\hat{p}_k(x), k = 1, \ldots, n\}$, i.e., it is not always true that $\hat{y}(x) = \arg\max\{\hat{p}_k(x), k = 1, \ldots, n\}$. It turns out that the two approaches, model prediction with type = "response", and model prediction via type = "probabilities", use different methods (see Wu et al., 2004, for details). For the examples considered in this paper, we implemented both approaches for predicting class labels, and we realized that the first approach with type = "response" yields more accurately predicted class labels than the latter approach with predicted class label as $\hat{y}(x) = \arg\max\{\hat{p}_k(x), k = 1, \ldots, n\}$. An illustration is given in the Results section.

Classification and regression tree (CART)

Classification trees gained much popularity in the machine learning literature since CART was developed by Breiman et al. (1984). In the context of image classification, the main idea is to come up with a decision tree that partitions the image via recursive partitioning into homogeneous regions, for instance, built-up, water, trees, and so on. We only discuss binary trees, as the decision trees with multiway splits can easily be obtained by iterative binary splits in binary trees.

The construction of a binary decision tree starts with assigning the entire training data D (N points) in one group called the root node. This node is then split into two nodes via one of the p predictors. For instance, X^i can be used to split the entire data into two subgroups or nodes $\{x : x^i \leq a\}$ and $\{x : x^i \leq a\}$. The two nodes are then further split using a value of another (or the same) predictor variable. The splitting process continues until a full tree is grown. Subsequently, techniques like cross-validation and tree complexity are used to prune the branches with very few data points to avoid over-fitting. Finally, each terminal (or leaf) node is assigned one class label $k \in \{1, 2, \ldots, n\}$

Choosing the best splitting variable and split point combination (X^i, a) at every node is an important part of the decision tree construction. An optimal (X^i, a) combination is obtained by minimizing the total within partition misclassification error $P_E(x) = 1–P_{\max}(x)$ (or some other impurity indices like Gini index or entropy) over every predictor-split point combination.

We used the implementation of CART in the R package "rpart" (Therneau et al., 2013) for all examples in this paper. First, rpart() is used to grow the full tree, then prune.rpart() prunes the tree using a tree complexity parameter (cp). Note that the value of cp has to be provided in rpart() as a control parameter; however, it can be tuned afterwards based on the cross validation error and the number of splits in the fully grown tree. Interestingly, for all examples considered in the paper, the optimal value of tree complexity parameter turns out to be cp = 0.01 (the default value). We used mostly the default parameters of rpart(), and prune.rpart(), except minsplit = 10 (in rpart.control), which determines the minimum number of points a node must have to be considered for splitting, and xval = 5 (in rpart.control), which specifies 5-fold cross-validation. As in ksum (), one can use predict.rpart() with type = "class" and type = "prob" to obtain the predicted class labels $\hat{y}(x)$ and multiclass classification probabilities $\{\hat{p}_k(x), k = 1, \ldots, n\}$. Unlike ksvm(), here, $\hat{y}(x) = \text{argmax}\{\hat{p}_k(x), k = 1, \ldots, n\}$. for all x in the input space (image).

Bayesian additive regression tree (BART)

Chipman et al. (2010) proposed a flexible non-parametric regression model called BART, which is a sum of trees model developed in the Bayesian framework. This paper

proposes a new multiclass classifier (referred to as mBACT) based on BART models. In this section, we present a brief overview of BART.

The BART model represents the response y at an input x as a sum of m adaptively chosen functions and an independent normal error,

$$y(x) = g(x; T_1, M_1) + g(x; T_2, M_2) + \ldots + g(x; T_m, M_m) + \varepsilon = h(x) + \varepsilon, \quad (2)$$

where the function g(x;T_j, M_j) denotes a "regression tree" model for a binary decision tree T_j with B_j terminal nodes (i.e., the input space is partitioned into B_j rectangular regions), $\varepsilon \sim N(0, \sigma^2)$, and $M_j = \{\mu_{j1}, \mu_{j2}, \ldots, \mu_{jB_j}, \}$ is the collection of terminal node predictions for tree T_j. As a function of x, the jth tree model $g(x;T_j, M_j)$ produces a piecewise-constant output, μ_{jb_j} which is obtained by following the sequence of decision rules in tree T_j and arriving at the terminal node $1 \leq b_j \leq B_j$. The model prediction at any input x is obtained by combining the ensemble of m trees output, i.e., $h(x) = \sum_{j=1}^{m} \mu_{jb_j}$. Viewing Equation (2) as a statistical model, the parameter vector $\Theta = (T_1, \ldots, T_m, M_1, \ldots, M_m, \sigma)$ has to be estimated. Chipman et al. (2010) take large values of m (50–200) and fit the model of Equation (2) in a Bayesian framework using Markov Chain Monte Carlo (MCMC) methods.

Each of the m decision trees is constructed using an extension of the Bayesian CART methodology (Chipman et al., 1998), where the basic idea is to specify a prior on the tree space and a prior on the terminal nodes outputs for each tree in the tree space. Instead of a closed-form prior, a tree-generating stochastic process was used on the tree space. Combining this prior with the tree model likelihood yields a posterior distribution on the tree space. An efficient MCMC-based algorithm was used to stochastically search for good trees in the tree space. As BART uses a sum of trees model, the trees with fewer splits (i.e., less than 5 splits) were assigned higher prior probabilities, and discrete uniform prior was used to choose the set of candidate split points for every split. For more details on these priors see Chipman et al. (2010).

We used the R library BayesTree (Chipman and McCulloch, 2009) for implementing BART (a key component of mBACT). The main function is bart(), which takes several arguments for controlling different features of the MCMC chains, and the priors on tree parameters, terminal node predictions, μ_{jb_j}, and the noise variance. The arguments related to the MCMC chain specifies that the predictive samples are saved every "keepevery" rounds after "nskip" samples are discarded as burn-in, and the chain stops after "ndpost" realizations. A few important tree parameters are "numcut" (the maximum number of split points along each input), "ntree" (the number of trees in the ensemble (m)), and "k" (the variance parameter in terminal node predictions), $\mu_{jb_j} \sim N(0, \sigma_\mu^2)$, where $\sigma_u \alpha 1/k\sqrt{m}$. Prior on the noise variance (σ^2) can also be passed in via "sigest", "sigdf" and "sigquant".

Compared with a single tree-based CART model, the BART model (Equation (2)) is an "ensemble" of m decision trees, and thus creates a flexible modeling framework. BART is also capable of incorporating higher dimensional interactions, by adaptively choosing the structure and individual rules of T_js. The sum of trees aspect of BART implicitly shares information from neighbouring inputs and models the spatial dependency. Furthermore, many individual trees (T_j) may place split points in the same area, allowing the predicted function to change rapidly nearby, effectively capturing nonstationary (spiky) behaviour such as abrupt changes in the response (e.g., between roads, grassland and water).

BART as a binary classifier

Like many other statistical regression models, BART can also be used for classification. Chipman et al. (2010) proposed an extension of BART called "BART probit" for classifying binary response $y \in \{0, 1\}$. The main idea is to model the response as

$$p(x) = P(Y = 1|x) = \Phi[h(x)]$$

where $h(x) = \sum_{j=1}^{m} g(x; T_j, M_j)$ and Φ is the standard normal cumulative distribution function (CDF), corresponding to a probit link function. Let D_0 be a training dataset with binary response $y_i \in \{0, 1\}$ at p-dimensional input $x_i = (x_i^1, x_i^2, \ldots, x_i^p)$, for $i = 1, \ldots, N$. For model convenience the response is rescaled to [−3, 3]. The BART probit model fit returns the posterior realizations of T_js and M_js, which are used to compute Monte Carlo estimate of $p(x)$. If the posterior draws from MCMC runs are $(T_j^{(s)}, M_j^{(s)}), j = 1, \ldots, m; s = 1, 2, \ldots, S$, the trained classifier $\hat{y}(x)$ is given by

$$\hat{y}(x) \begin{cases} 1, & if \ \frac{1}{S} \sum_{s=1}^{S} \Phi[h^s(x)] \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $h^{(s)}(x) = \sum_{j=1}^{m} g\left(x; T_j^{(s)}, M_j^{(s)}\right)$ and $\hat{p}(x) = \frac{1}{S} \sum_{s=1}^{S} \Phi\left[h^{(s)}(x)\right]$.

From an implementation viewpoint, the built-in function bart() in the R library BayesTree can also be used to fit BART probit models. It is however worth noting that for BART probit implementation a few arguments of bart() are now either ignored by the function or have slightly different values. For instance, in case of binary response, the noise variance (σ^2) is fixed at 1, and as a result "sigest", "sigdf", and "sigquant" are ignored. The prior variance on the terminal node output is defined by $\sigma_u = 3/k\sqrt{m}$ compared with $\sigma_u = 0.5/k\sqrt{m}$ (used in the regression framework). One can also include an argument called "binaryOffset" for offsetting the value of $h(x)$ away from zero.

Chipman et al. (2010) used BART probit model in a drug discovery application, where the goal was to classify compounds. Zhang and Hardle (2010) independently developed a similar adaptation of BART, applying it to the problem of classifying the solvency status of German firms based on their financial statement information. Zhang and Hardle called their classifier Bayesian Additive Classification Trees (BACT) and demonstrated that BACT outperforms CART and SVM in identifying insolvent firms.

New methodology: mBACT

In this section we propose a new reliable classifier called mBACT, which is a multiclass generalization of BART probit and BACT. The key idea is to use BART probit with the one-against–all approach for developing a multiclass classifier. Though the motivating application considered in this paper comes from remote sensing literature, the classifier proposed here can be used in other applications as well.

Let D be a set of N points with response $y_i \in \{1,2,\ldots, n\}$ and p-dimensional input $x_i = (x_i^1, x_i^2, \ldots, x_i^p)$, for $i = 1,\ldots, N$. The method starts with transforming the data D that would facilitate the one-against-all approach. For each class $k = 1, \ldots, n$, we generate a pseudo data set D_k with original x_i, and new response $y_{i(k)}$ defined as follows:

$$y_{i(k)} \begin{cases} 1, & if\ y_i = k \\ 0, & f\ y_i \neq k \end{cases} \tag{4}$$

In the spirit of Liu and Zheng (2005), we build n binary classifiers using D_1, D_2, …, D_n, and then combine them to obtain the desired multiclass classifier. For $k = 1,2, \ldots, n$, let $p_{(k)}(x) = p(Y_{(k)} = 1|x)$ be the classification probability of input x with class label 1 (indicator of class k against others) under the binary data $D_k = \left\{(x_i, y_{i(k)}), i = 1, \ldots, N\right\}$ (note that $p_{(k)}(x)$ and $p_{(k)}(x)$ are slightly different quantities). Then, the data set D_k is used to fit the standard BART probit model

$$p_{(k)}(x) = p\left(y_{(k)} = 1|x\right) = \Phi\left[h_{(k)}(x)\right]$$

where $h_{(k)}(x) = \sum_{j=1}^{m} g_{(k)}\left(x; T_{j(k)}, M_{j(k)}\right)$ and Φ is the standard normal CDF. The Monte Carlo estimate of $p_{(k)}(x)$ obtained from the posterior draws of $T_{j(k)}$s and $M_{j(k)}$s is

$$\hat{p}_{(k)}(x) = \frac{1}{S}\sum_{s=1}^{S} \Phi\left[h_{(k)}^{(s)}(x)\right]$$

The classification probabilities of these n binary classifiers based on D_1, D_2, …, D_n can be combined to obtain the predicted class label under the original n-class dataset D,

$$\hat{y}(x) = \arg\ max\left\{\hat{p}_{(k)}(x), k = 1, \ldots, n\right\} \tag{5}$$

Because $p_{(k)}(x)$ are binomial success probabilities for n data sets and not multinomial probabilities for one data set, the total $\sum_{k=1}^{n} \hat{p}_{(k)}(x)$ is not always 1. However, $\hat{p}_{(k)}(x)$ can be approximated using $\hat{p}_{(k)}(x)$ as follows

$$\hat{p}_{(k)}(x) \approx \frac{\hat{p}_{(k)}(x)}{\sum_{l=1}^{n} \hat{p}_{(l)}(x)}, \text{for}\, k = 1, \ldots, n \tag{6}$$

From Equations (5) and (6), it is clear that the predicted class label matches with the maximum predicted classification probability $\hat{p}_{\max}(x) = \max\{\hat{p}_k(x), k = 1, \ldots, n\}$.

The implementation of mBACT requires fitting BART probit models on n pseudo data sets D_1, D_2, …, D_n. For all examples considered in this paper, we used mostly the default arguments of bart(), except different MCMC parameters, keepevery = 20, ndpost = 5000 (for more stable posterior estimates), tree parameter, numcut = 1000 (for refined search of optimal split points), and variance parameter $k = 1$. The default value of $k = 2$ imposes considerable shrinkage to the individual terminal node output μ_{jb_j}, whereas the proposed change ($k = 1$) increases the prior variance and applies less shrinkage (or smoothness) of the response. This is particularly important for our application as the land use (e.g., the buildup area and water bodies) changes abruptly.

Study area and data collection

A multispectral satellite image acquired by LANDSAT 5 TM on 15 August 2009 (path–row: 08–29, L1T product, UTM map projection) over Nova Scotia, Canada, is considered for this comparative study. Although LANDSAT 5 TM data consist of seven bands, the sixth band is thermal with coarser resolution (120 m) than the other six reflectance bands (30 m), thus we used only the six reflectance bands (blue, green, red, near infrared, and two middle infrared bands).

This LANDSAT scene (185 km × 170 km) covers (43.632 N, 63.266 W) to (45.579 N, 65.169 W), where each pixel is of 30 m × 30 m resolution. Based on the consistency of land use and accessibility of ground data, we chose three relatively small regions of Kings County (the towns of Wolfville, Windsor, and Kentville and their surrounding rural areas) from this scene for performance comparison.

(*i*) Wolfville area: A small portion of the scene covering (45.070 N, 64.334 W) to (45.098 N, 64.386 W), with 105 × 134 pixel image (**Figure 1b**).

(*ii*) Windsor area: A medium size region of the scene covering (44.932 N, 64.102 W) to (44.995 N, 64.195 W) with 236 × 239 pixels (**Figure 1c**).

(*iii*) Kentville area: A relative large region of the scene from (45.044 N, 64.418 W) to (47.117 N, 64.552 W) with 278 × 349 pixels (**Figure 1d**).

A false color composite of the three study areas constructed using three bands (green, red, and near infrared) are shown in **Figure 1**.

Each pixel of the image in the three study areas can be classified into one of the seven classes: built-up, pond–lake–river water, Bay of Fundy, agricultural land or barren, grassland, trees and scrubland. The land cover of Windsor is interesting as it is clear from the map (**Figure 1a**) that the water body passing from the west to the north side of **Figure 1c** is a part of the Bay of Fundy that is cutoff due to the construction of a causeway for Highway 101. Because this water is not tidal, the "Bay of Fundy" class is not used in the Windsor scene.

The detailed class-by-class breakdown of the training and validation data size of the three study areas are shown in **Table 1**. The data were collected by sampling several disjointed homogeneous patches without replacement using class-wise stratified random sampling. Note that the sample size increases with the size of the region.

For all three study areas, we followed the same approach of distributing the data points in the training and validation sets, that is, for each class, approximately two-thirds of the data points were assigned to the training set and the remainder to the validation set.

The data points (i.e., the ground truth or true class labels) were collected using a combination of in-person site visits, Google street views, and Google satellite views. Although the data points were collected in 2012–2013, the land use has not changed much (except a few differences in the built-up and scrubland classes) since 2009 (when the satellite image was taken). We want to emphasize that the main purpose of this paper was to compare the classification accuracy of mBACT with SVM and CART, which should not be affected by a few incorrectly labeled ground observations.

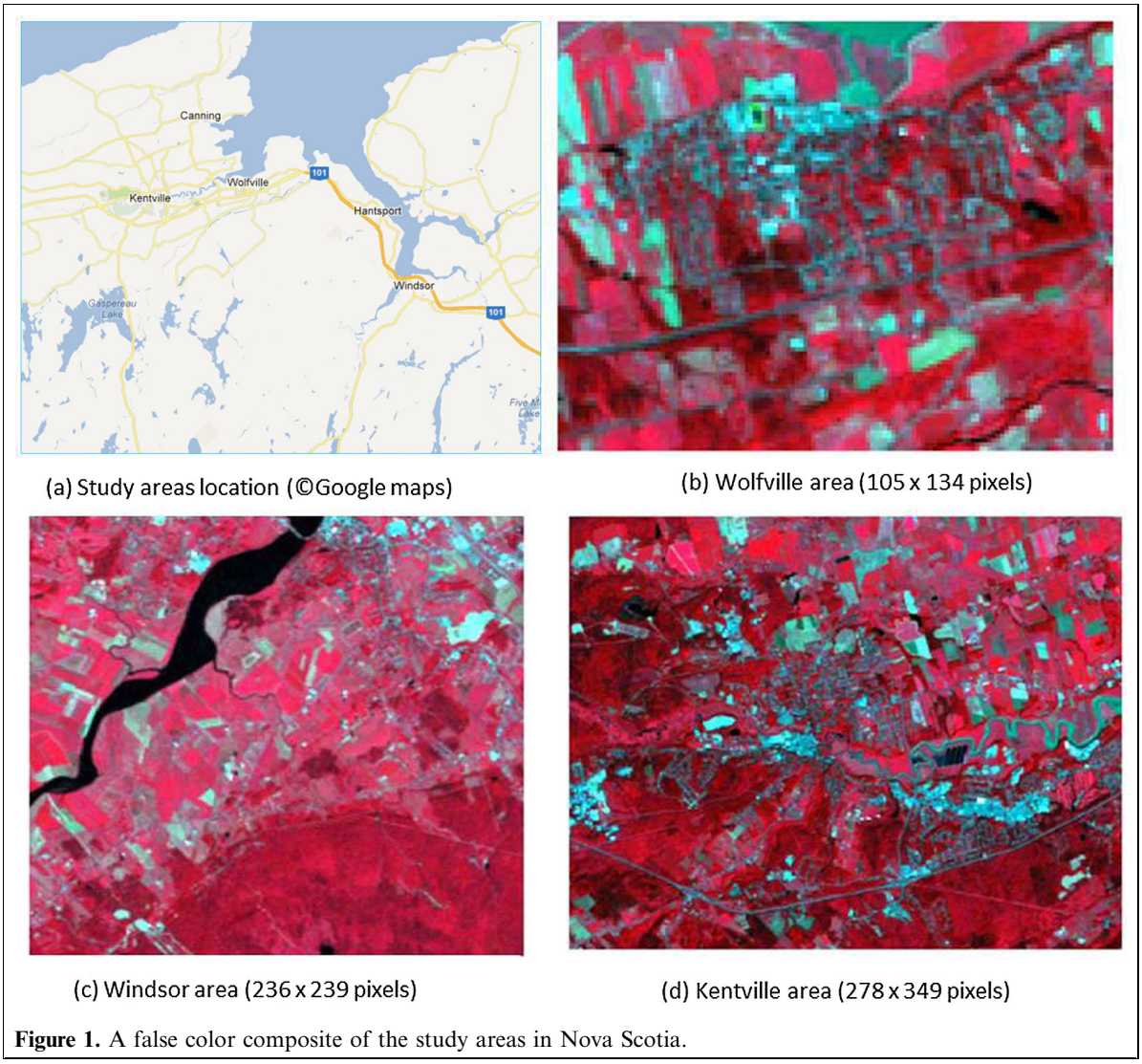

Figure 1. A false color composite of the study areas in Nova Scotia.

Table 1. Class-by-class distribution of the training and validation samples for each of seven classes in the three study areas.

	Wolfville			Windsor			Kentville		
Classes	Training	Validation	Total	Training	Validation	Total	Training	Validation	Total
Built-up	37	17	54	76	42	118	188	96	284
Water	12	10	22	27	13	40	35	19	54
Bay of Fundy	14	12	26	0	0	0	34	18	52
Agricultural land	31	14	45	27	17	44	78	46	124
Grassland	21	17	38	61	31	92	81	41	122
Trees	16	10	26	61	32	93	112	57	169
Scrubland	23	15	38	69	33	102	46	23	69
Total	154	95	249	321	168	489	574	300	874

Results and discussion

Accuracy measures

For a given classifier and a study area, let $F = ((f_{ij}))$ be the error (or confusion) matrix, where f_{ij} denotes the number of validation or reference points in the jth class with predicted class labels i. Then, f_{ii} is the number of correctly classified validation points in the ith class, f_{i+} (ith row sum) is the number of validation points predicted to be in class i, and f_{i+} (ith column sum) is the true number of validation points in class i. Define $\eta_1 = \left(\sum_{i=1}^{n} f_{ii}\right)/V, \eta_2 = \left(\sum_{i=1}^{n} f_{i+}f_{+i}\right)/V^2, \eta_3 = \left[\sum_{i=1}^{n} f_{ii}(f_{i+}+f_{+i})\right]/V^2$, and $\eta_4 = \left[\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}\left(f_{+i}+f_{j+}\right)^2\right]/V^3$ Where $V = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{n} f_{+i} = \sum_{i=1}^{n} f_{i+}$, is the grand total of the error matrix or the size of the validation data. Then, η_1 measures the overall accuracy of predicted class labels. Class-wise accuracies from user's and producer's perspectives can be measured by f_{ii}/f_{i+} and f_{ii}/f_{+i} respectively. The remaining quantities η_2, η_3, and η_4 are used in defining another popular accuracy measure called kappa (κ) which quantifies the agreement between the predicted class labels and the reality. The overall kappa coefficient is estimated by

$$\hat{k} = \frac{\eta_1 - \eta_2}{1 - \eta_2}$$

and the associated uncertainty is measured by

$$\mathrm{var}(\hat{\kappa}) = \frac{1}{V}\left\{\frac{\eta_1(1-\eta_1)}{(1-\eta_2)^2} + \frac{2(1-\eta_1)(2\eta_1\eta_2 - \eta_3)}{(1-\eta_2)^3} + \frac{(1-\eta_1)^2(\eta_4 - 4\eta_2^2)}{(1-\eta_2)^4}\right\}$$

The prediction accuracy for ith class can be measured by conditional kappa

$$\hat{\kappa}_I = \frac{Vf_{ii} - f_{i+}f_{+i}}{Vf_{i+} - f_{i+}f_{+i}}$$

with the associated variance

$$\mathrm{var}(\hat{\kappa}_I) = \frac{V(f_{i+} - f_{ii})}{\left[f_{i+}(V - f_{+i})\right]^3} \times\{(f_{i+} - f_{ii})(f_{i+}f_{+i} - Vf_{ii}) + Vf_{ii}(V - f_{+i} - f_{i+} + f_{ii})\}$$

Note that all of these accuracy measures are based on the discrepancy between the predicted and true class labels. Next, we present a few uncertainty measures that are based on the multiclass classification probabilities.

Uncertainty measures

Assuming $\{\hat{p}_k(x), k = 1, \ldots, n\}$ are estimated multinomial probabilities for the class labels of any pixel (or site) x in the validation data, one can define the deviance as

$$D = -2\left(\sum_{i=1}^{V}\sum_{k=1}^{n} \log(\hat{p}_k(x_i)) \cdot \mathrm{I}[\hat{y}(x_i) = k]\right)$$

where $\mathrm{I}[\hat{y}(x_i) = k] = 1$, if the predicted class label is k, and zero otherwise. Small values of deviance represent confident prediction of correct class labels (i.e., less uncertainty). Because $\hat{y}(x_i)$ corresponds to $\max\{\hat{p}_k(x), k = 1, \ldots, n\}$ for all x, in both CART and mBACT and not in SVM, it is expected that the deviances for CART and mBACT would be smaller compared with that for SVM classified images.

We also used a few $p_k(x)$ based impurity indices for measuring uncertainty in predicting the class labels for every input site x in the study area.

- Probability of miss-classification: $P_E(x) = 1 - \max\{p_k(x), k = 1, 2, \ldots, n\}$
- Gini index: $G(x) = 1 + \sum_{k=1}^{n}[p_k(x)]^2$
- Entropy: $H(x) = -\sum_{k=1}^{n} p_k(x)\log(p_k(x))$

In practice, we used the estimates of $p_k(x)$ for computing these impurity indices. Because none of these three measures are linked with the predicted class, we do not expect a systematic bias towards a particular classifier.

Classified images and tabulated results

We compared the classified images obtained from SVM, CART, and mBACT. Both overall and class-wise goodness-of-fit measures were tabulated and compared for these classifiers. Because the impurity indices (P_E (x), $G(x)$, and H (x)) were computed for every site in the study area, one could compare the uncertainty images instead of class-wise averages; however, due to limited space we only present uncertainty images for Kentville area. The results presented

here are ordered based on the size of the region: Wolfville, Windsor, and Kentville. **Figure 2** shows the classified images of Wolfville area. The predicted class labels for "built-up" and "trees" in CART appear to be relatively noisier compared with that obtained from SVM and mBACT. A quick comparison of SVM with mBACT generated images does not show much difference. However, a closer look at the accuracy measures (**Table 2**) reveals more precise information and a clear overall trend in the prediction accuracy and uncertainty.

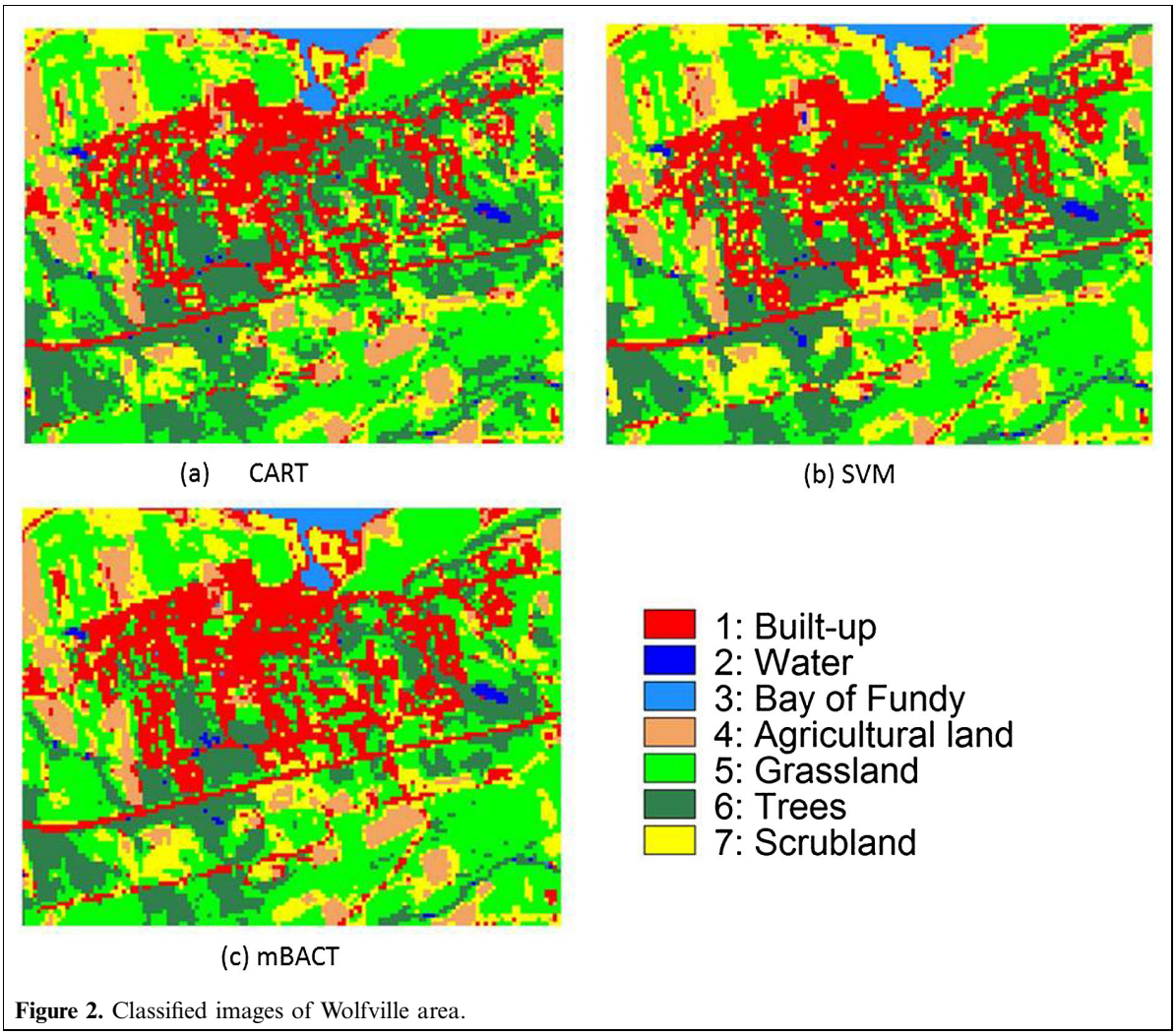

Figure 2. Classified images of Wolfville area.

Table 2. Overall and class-wise accuracy and uncertainty measures for Wolfville.

	CART-Wolfville*					SVM-Wolfville†					mBACT-Wolfville‡				
Classes	User's	Producer's	Co. kappa	Gini	Entropy	User's	Producer's	Co. kappa	Gini	Entropy	User's	Producer's	Co. kappa	Gini	Entropy
Built-up	80.95	100.00	0.768	0.000	0.000	94.12	94.12	0.928	0.535	1.172	80.00	94.12	0.756	0.558	0.917
Water	100.00	80.00	1.000	0.000	0.000	90.00	90.00	0.888	0.531	1.147	100.00	80.00	1.000	0.501	0.933
Bay of Fundy	100.00	75.00	1.000	0.000	0.000	100.00	91.67	1.000	0.685	1.485	100.00	83.33	1.000	0.282	0.758
Agricultural land	59.09	92.86	0.520	0.233	0.462	66.67	100.00	0.609	0.763	1.653	66.67	100.00	0.609	0.372	0.718
Grassland	100.00	94.12	1.000	0.135	0.211	94.12	94.12	0.928	0.516	1.131	89.47	100.00	0.872	0.385	0.792
Trees	76.92	100.00	0.742	0.111	0.224	100.00	90.00	1.000	0.626	1.375	90.91	100.00	0.898	0.536	0.901
Scrubland	100.00	40.00	1.000	0.234	0.470	90.00	60.00	0.881	0.771	1.684	83.33	33.33	0.802	0.566	0.963

*Overall η_1 = 83.16%; kappa (κ) = 0.802; var(kappa) = 0.0020; overall Gini = 0.127; overall entropy = 0.236; overall deviance = 12.03.
†Overall η_1 = 88.42%; kappa (κ) = 0.864; var(kappa) = 0.0015; overall Gini = 0.599; overall entropy = 1.310; overall deviance = 110.69.
‡Overall η_1 = 84.21%; kappa (κ) = 0.814; var(kappa) = 0.0019, overall Gini = 0.471; overall entropy = 0.856; overall deviance = 48.62.

Table 2 shows that in terms of overall accuracy measures (η_1 and κ), SVM generates the most accurate predicted class label in the validation set. Furthermore, mBACT is better than CART but inferior to SVM (i.e., SVM > mBACT > CART). Class-wise accuracy measures (user's and producer's accuracy and condition kappa) do not exhibit a consistent trend over the seven classes.

In terms of overall uncertainty measures (deviance, Gini, and entropy), **Table 2** exhibits a different overall trend (CART > mBACT > SVM). Unlike accuracy measures, class-wise uncertainty measures (Gini and entropy) show the same consistent trend (CART > mBACT > SVM) for every class as well. By combining the information on accuracy and uncertainty measures, it appears that CART is a more confident but less accurate predictor than mBACT (mBACT > CART in terms of η_1 and κ). Because the predicted class labels do not necessarily match with the maximum predicted classification probabilities, it is not surprising that SVM leads to the largest deviance. However, the ranking of SVM based on overall and class-wise Gini and entropy values indicate that SVM-based predictions are more uncertain than mBACT.

The classified images of Windsor area are shown in **Figure 3**. A quick view of **Figure 3** shows that all three classifiers perform reasonably well in capturing the main features of the land-use. See **Table 3** for a detailed performance comparison of the classifiers.

In terms of both overall accuracy measures (η_1 and κ), it is clear from **Table 3** that mBACT is slightly better than SVM and much better than CART. In this case, even the class-wise accuracy measures (user's, producer's, and conditional kappa) support the superior performance of mBACT in most of the classes. The uncertainty measures (deviance, Gini, and entropy) follow the same trend as in the Wolfville image.

Figure 4 displays the classified images of Kentville area. It is clear from **Figure 4** that CART yields a somewhat noisier classified map (particularly in built-up and scrubland classes) than SVM and mBACT.

Goodness-of-fit measures for the Kentville area images are summarized in **Table 4**. The overall accuracies (η_1 and κ) indicate that mBACT is comparable with SVM and produces more accurate predicted class labels than CART.

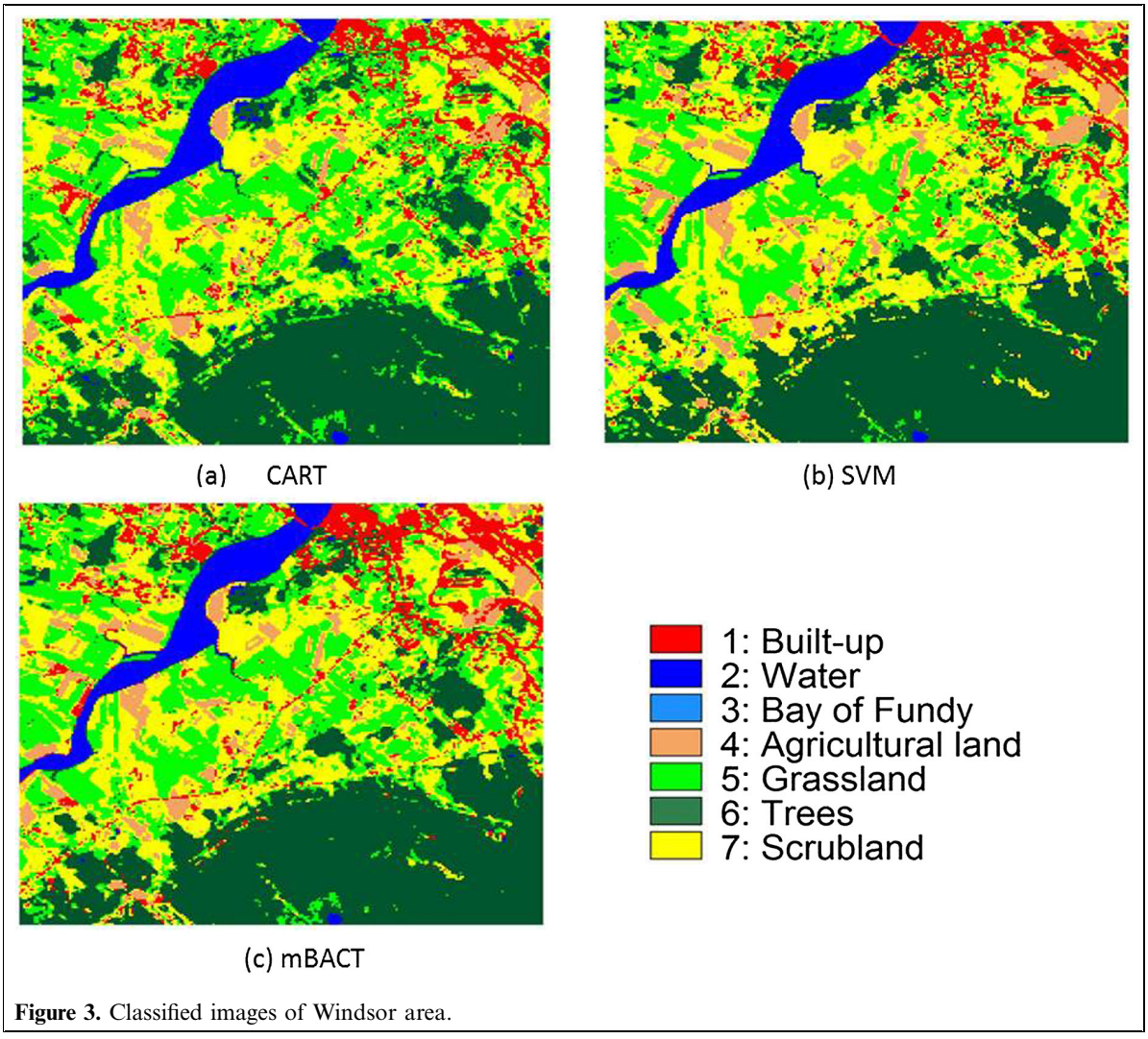

Figure 3. Classified images of Windsor area.

Table 3. Overall and class-wise measures of accuracy and uncertainty for Windsor.

	CART-Windsor*					SVM-Windsor†					mBACT-Windsor‡				
Classes	User's	Producer's	Co. kappa	Gini	Entropy	User's	Producer's	Co. kappa	Gini	Entropy	User's	Producer's	Co. kappa	Gini	Entropy
Built-up	97.06	78.57	0.961	0.182	0.363	95.00	90.48	0.933	0.446	0.903	95.00	90.48	0.933	0.413	0.603
Water	86.67	100.00	0.856	0.069	0.154	100.00	100.00	1.000	0.117	0.254	100.00	100.00	1.000	0.107	0.385
Bay of Fundy	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Agricultural land	87.50	82.35	0.861	0.180	0.325	100.00	76.47	1.000	0.594	1.249	93.75	88.24	0.931	0.364	0.648
Grassland	80.00	90.32	0.755	0.191	0.335	90.62	93.55	0.885	0.491	1.058	93.33	90.32	0.918	0.362	0.534
Trees	93.10	84.38	0.915	0.062	0.143	87.88	90.62	0.850	0.350	0.773	93.75	93.75	0.923	0.201	0.333
Scrubland	82.05	96.97	0.775	0.134	0.329	89.19	100.00	0.866	0.592	1.250	89.19	100.00	0.866	0.389	0.598

*Overall η_1 = 87.50%; kappa (κ) = 0.847; var(kappa) = 0.0010; overall Gini = 0.130; overall entropy = 0.268; overall deviance = 22.99.
†Overall η_1 = 92.26%; kappa (κ) = 0.905; var(kappa) = 0.0006; overall Gini = 0.463; overall entropy = 0.990; overall deviance = 129.30.
‡Overall η_1 = 93.45%; kappa (κ) = 0.919; var(kappa) = 0.0005; overall Gini = 0.311; overall entropy = 0.492; overall deviance = 49.73.

Similar to the Wolfville image, the class-wise accuracy measures (user's, producer's, and conditional kappa) do not exhibit a clear trend across the classifiers, but the uncertainty measures show a consistent pattern with CART being the most confident classifier and SVM the most uncertain.

Note that the value of deviance has been increasing with the size of the validation set, because it is a sum and not an average.

Figure 5 presents site-wise comparison of the probability of misclassification $p_E(x)$, Gini index $G(x)$, and entropy $H(x)$ for all three classifiers. All uncertainty images support

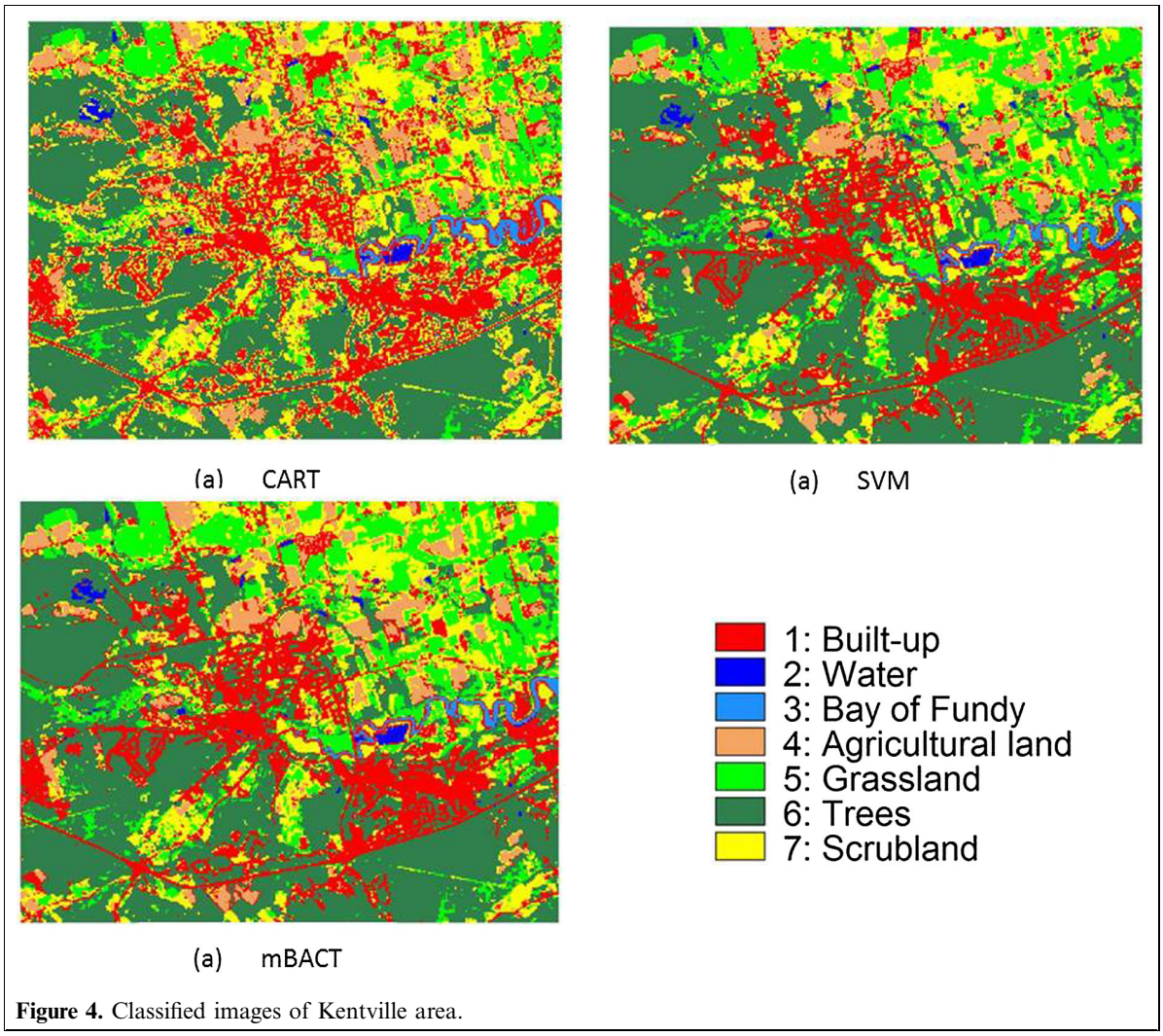

Figure 4. Classified images of Kentville area.

Table 4. Overall and class-wise measures of accuracy and uncertainty for Kentville.

	CART-Kentville*					SVM-Kentville†					mBACT-Kentville‡				
Classes	User's	Producer's	Co. kappa	Gini	Entropy	User's	Producer's	Co. kappa	Gini	Entropy	User's	Producer's	Co. kappa	Gini	Entropy
Built-up	83.33	83.33	0.755	0.126	0.281	89.11	93.75	0.840	0.507	1.147	87.38	93.75	0.814	0.421	0.671
Water	95.00	100.00	0.947	0.059	0.136	89.47	89.47	0.888	0.444	0.948	86.36	100.00	0.854	0.359	0.601
Bay of Fundy	82.35	77.78	0.812	0.171	0.405	100.00	83.33	1.000	0.721	1.577	100.00	72.22	1.000	0.258	0.619
Agricultural land	84.85	60.87	0.821	0.353	0.665	92.31	78.26	0.909	0.778	1.726	90.00	78.26	0.882	0.435	0.715
Grassland	90.24	90.24	0.887	0.106	0.211	88.10	90.24	0.862	0.725	1.551	88.37	92.68	0.865	0.384	0.649
Trees	100.00	86.49	1.000	0.036	0.103	91.80	98.25	0.899	0.783	1.723	100.00	96.49	1.000	0.245	0.460
Scrubland	52.63	86.96	0.487	0.517	1.011	82.61	82.61	0.812	0.776	1.676	79.17	82.61	0.774	0.564	0.902

*Overall η_1 = 84.33%; kappa (κ) = 0.807; var(kappa) = 0.0007; overall Gini = 0.204; overall entropy = 0.418; overall deviance = 50.29.
†Overall η_1 = 90.00%; kappa (κ) = 0.875; var(kappa) = 0.0005; overall Gini = 0.716; overall entropy = 1.571; overall deviance = 393.15.
‡Overall η_1 = 90.00%; kappa (κ) = 0.875; var(kappa) = 0.0006; overall Gini = 0.365; overall entropy = 0.619; overall deviance = 106.67.

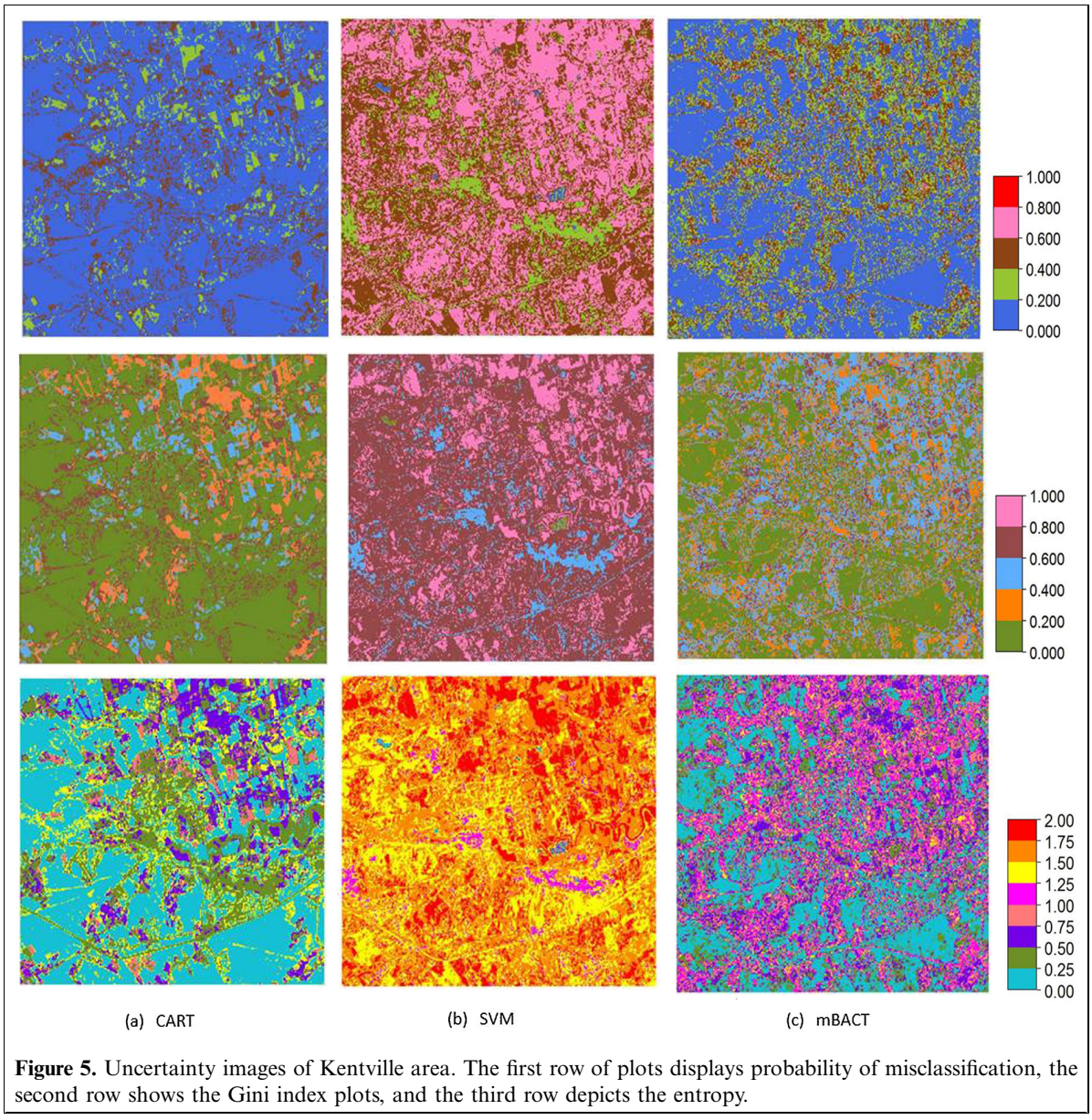

Figure 5. Uncertainty images of Kentville area. The first row of plots displays probability of misclassification, the second row shows the Gini index plots, and the third row depicts the entropy.

the expected trend, that is, CART is the most confident classifier, and SVM-based predictions are most uncertain.

Overall summary

The performance comparison of mBACT with SVM and CART based on the three study areas suggest the following. (*i*) Based on the overall accuracy measures (η_1 and κ), mBACT outperforms CART in all cases and performs better than SVM in one of three cases. In one case (Wolfville), SVM yielded higher overall accuracy than mBACT, and in one case (Kentville) SVM and mBACT are identical. (*ii*) In terms of uncertainty measures (deviance, Gini and entropy), mBACT generates predictions with slightly larger $\hat{p}_{\max}(x)$ compared with SVM, and CART turns out to be the most confident predictor.

Given that CART yields the smallest values of η_1 and κ for all three images, the overconfidence in terms of deviance, Gini, and entropy is somewhat questionable. To investigate this further we present a display that evaluates the accuracy of prediction of the class probabilities. This is accomplished by comparing the maximum predicted probability $\hat{p}_{\max}(x)$ with the actual class label Y for each observation in the validation set. A direct comparison between $\hat{p}_{\max}(x)$ and Y at the level of individual observations is not practical, because the observed class label will either equal the class that has maximum predicted probability or it will not. In other words, such a comparison would be between a predicted probability and a binary indicator for whether the observed class is the same as the predicted class. However, such a comparison can be made by combining observations into groups and comparing the proportion of correctly predicted classes to the average value of $\hat{p}_{\max}(x)$ in each group. We group the observations as follows: for each observation obtain $\hat{p}_{\max}(x)$; then sort the observations from smallest to largest value of $\hat{p}_{\max}(x)$; divide these sorted observations into 10 groups; thus the 1st group will consist of observations with the smallest amount of certainty in

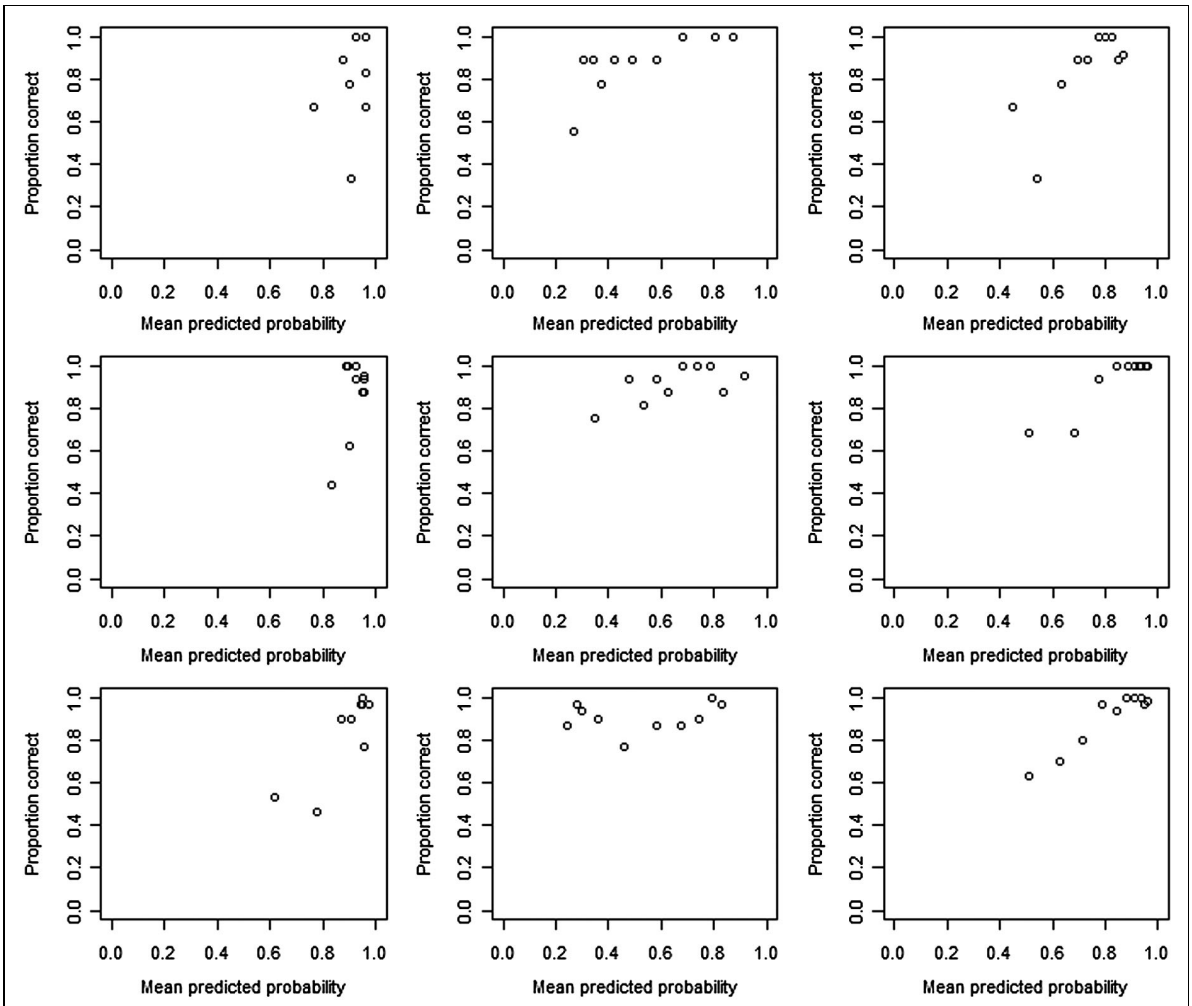

Figure 6. Each plot presents "proportion of correctly classified validation points" versus "average $\hat{p}_{\max}(x)$" in 10 $\hat{p}_{\max}(x)$ quantile bins over the validation set. (1st row, Wolfville; 2nd row, Windsor; 3rd row, Kentville; 1st column, CART; 2nd column, SVM; 3rd column, mBACT).

prediction, and the 10th group will consist of observations with the largest prediction certainty. For each group, also calculate the proportion of observations for which the class with maximum predicted probability was equal to the observed class. If a model is accurately predicting the class probabilities, then this proportion of correctly classified observations in each group should equal the mean of the maximum predicted class probabilities. A plot of these quantities for the 10 groups should correspond to a line with intercept 0 and slope 1 for the ideal predictor. **Figure 6** presents a comparison of such plot for all three classifiers in the three study areas.

It is clear from **Figure 6** that mBACT (the third column) is the most reliable classifier in this measure (with a slope closest to 1), and CART (the first column) leads to the most confident prediction (with points whose mean predicted $\hat{p}_{\max}(x)$ exceed the proportion correctly classified within each group). Note that in mBACT-Wolfville plot, the smallest ''proportion of correct classification'' point with 30% correct classification corresponds to average $\hat{p}_{\max}(x) \approx 0.55$, whereas in CART-Wolfville, average $\hat{p}_{\max}(x) \approx 0.9$. As a result CART can be an unreliable (i.e., overconfident) classifier.

Although we include plots in **Figure 6** for SVM, we note that the SVM package uses two different models to make a class prediction and to predict class probabilities. Thus there is less reason to expect that the points in the middle column of **Figure 6** will have a line with slope 1, because the class labels used in calculating the proportion of correct classifications (vertical axis) are from a different model than the predicted class probabilities. Indeed, there seems to be a very weak correspondence between $\hat{p}_{\max}(x)$ and the proportion of correct classifications for SVM in **Figure 6**.

Concluding remarks

Accurate prediction of class labels of a satellite image has been a challenging problem in remote sensing applications. In this article, we introduced a new reliable multiclass classifier, mBACT, for accurate identification of class labels. Based on a small case study shown in this paper, it appears that mBACT clearly outperforms CART; however, SVM is almost on par with mBACT in terms of accuracy measures. Furthermore, the accuracy of SVM tends to exceed its prediction accuracy. Though the classification problem considered in this paper comes from a remote sensing application, mBACT can be used for other applications as well.

The main idea of mBACT was to generalize the binary classifier (BART probit or BACT) for the multiclass classification problem using the one-against-all approach. This requires fitting BART model n times (the number of classes) for the entire data set, and the current version of the *R* library BayesTree (for fitting BART models) is computationally more expensive than SVM and CART. This is somewhat expected because BART is a Bayesian ensemble of trees model, and CART is based on a single decision tree. Pratola et al. (2014) demonstrated that BART can be speeded up and applied to large datasets through the use of parallel computation methods. Further gains in multiclass problems might be realized by a one-against-one generalization of BART probit.

Acknowledgements

We would like to thank the referees for many useful comments and suggestions that led to significant improvement of the article. This work was supported in part by Discovery grants from the Natural Sciences and Engineering Research Council of Canada.

References

Anderson, J.R., Hardy, E.E., Roach, J.T., and Witmer, R.E. 1976. A land use and land cover classification system for use with remote sensor data. *U. S. Geological Survey Professional Paper, No. 964*, USGS, Washington, D.C.

Bazi, Y., and Melgani, F. 2006. Toward an optimal SVM classification system for hyperspectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 44, No. 11, pp. 3374–3385.

Bottou, L., Cortes, C., Denker, J., Drucker, H., Guyon, I., Jackel, L., LeCun, Y., Muller, U., Sackinger, E., Simard, P., and Vapnik, V. 1994. Comparison of classifiers methods—a case study in handwriting digit recognition. In *Proc. International Conference on Pattern Recognition*, pp. 77–87.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and regression trees.* Chapman and Hall/CRC, Boca Raton, Florida.

Chipman, H., George, E., and McCulloch, R. 1998. Bayesian CART model search. *Journal of the American Statistical Association*, Vol. 93, No. 443, pp. 935–948.

Chipman, H., and McCulloch, R. 2009. *BayesTree: Bayesian Methods for Tree Based Models.* R package version 0.3-1.1, URL: http://cran.r-project.org/web/packages/BayesTree.

Chipman, H.A., George, E.I., and McCulloch, R.E. 2010. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, Vol. 4, No. 1, pp. 266–298.

Franklin, S.E., Peddle, D.R., Dechka, J.A., and Stenhouse, G.B. 2002. Evidential reasoning with Landsat TM, DEM, and GIS data for land cover classification in support of grizzly bear habitat mapping. *International Journal of Remote Sensing*, Vol. 23, pp. 4633–4652.

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

Friedman, J. 1996. *Another approach to polychoto-mous classification* (Technical Report). Stanford University, Department of Statistics.

Gallego, F.J. 2004. Remote sensing and land cover area estimation. *International Journal of Remote Sensing*, Vol. 25, pp. 3019–3047.

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

Hsu, C.W., and Lin, C.J. 2002. A comparison of methods for multi-class support vector machines. *IEEE transactions on Neural Networks*, Vol. 13, No. 2, pp. 415–425.

Jensen, J.R. 1996. *Introductory digital image processing: a remote sensing perspective*. Pearson Prentice Hall, New Jersey, USA.

Karatzoglou, A., Smola, A., and Hornik, K. 2013. *Kernlab: Kernel-based Machine Learning Lab*. R package version 0.9-18, URL: http://cran.r-project.org/web/packages/kernlab.

Knerr, S., Personnaz, L., and Dreyfus, G. 1990. *Single-layer learning revisited: A stepwise procedure for building and training neural network. Neurocomputing: Algorithms, Architectures and Applications*, NATO ASI, Berlin: Springer-Verlag.

Lawrence, R., Bunn, A., Powell, S., and Zambon, M. 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, Vol. 90, pp. 331–336.

Li, M., Crawford, M.M., and Jinwen, T. 2010. Local manifold learning-based k-nearest neighbor for hyperspectral image classification. *IEEE Transactions on Geosciences and Remote Sensing*, Vol. 48, No. 11, pp. 4099–4109.

Liu, Y., and Zheng,Y.F. 2005. One-against-all multiclass SVM classification using reliability measures. In *Proc. IEEE International Joint Conference on Neural Networks*, Vol. 2, pp. 849–854.

Lu, D., and Weng, Q. 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, Vol. 28, No. 5, pp. 823–870.

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

Peddle, D.R. 1993. An empirical comparison of evidential reasoning, linear discriminant analysis and maximum likelihood algorithms for alpine land cover classification. *Canadian Journal of Remote Sensing*, Vol. 19, No. 1, pp. 31–44.

Pratola, M.T., McCulloch, R., Gattiker, J., Chipman, H., and Higdon, D. 2014. Parallel Bayesian additive regression trees, (in press). *Journal of Computational and Graphical Statistics*.

Song, X., Duan, Z., and Jiang, X. 2012. Comparison of artificial neural network and support vector machine classifiers for land cover classification in Northern China using a SPOT-5 HRG image. *International Journal of Remote Sensing*, Vol. 33, No. 10, pp. 3301–3320.

Sudha, L.R., and Bhavani, R. 2012. Performance comparison of SVM and k-NN in automatic classification of human gait pattern. *International Journal of Computers*, Vol. 6, No. 1, pp. 19–28.

Therneau, T., Atkinson, B., and Ripley, B. 2013. *Rpart: recursive partitioning and regression trees*, R package version 4.1-1, URL: http://cran.r-project.org/web/packages/rpart

Vapnik, V.B. 1995. *The Nature of Statistical Learning Theory*. Springer Verlag, New York.

Vapnik, V.B., Golowich, S.E., and Smola, A.J. 1996. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, Vol. 9, pp. 281–287.

Wacker, A.G., and Landgrebe, D.A. 1972. Minimum distance classification in remote sensing. *LARS Technical reports, Paper 25*. http://docs.lib.purdue.edu/larstech/25.

Wu, T.-F., Lin, C.-J., and Weng, R.C. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, Vol. 5, pp. 975–1005.

Yang, C.C., Prasher, S.O., Enright, P., Madramootoo, C., Burgess, M., Goel, P.K., and Callum, I. 2003. Application of decision tree technology for image classification using remote sensing data. *Agricultural Systems*, Vol. 76, No. 3, pp. 1101–1117.

Zhang, J.L., and Hardle, W.K. 2010. The Bayesian additive classification tree applied to credit risk modeling. *Computational Statistics and Data Analysis*, Vol. 54, pp. 1197–1205.

Zhang, Q., and Wang, J. 2003. A rule-based urban land use inferring method for fine resolution multispectral imagery. *Canadian Journal of Remote Sensing*, Vol. 29, pp. 1–13.