Remote Sensing of Environment 141 (2014) 14–23

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery

Linlin Xu [a], Jonathan Li [a,b,*], Alexander Brenning [a]

[a] University of Waterloo, Department of Geography and Environmental Management, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada
[b] Xiamen University, Key Laboratory for Underwater Acoustic Communication and Marine Information Technology, School of Information Science and Engineering, 422 Siming Road South, Xiamen, Fujian 361005, China

ARTICLE INFO

Article history:
Received 18 March 2013
Received in revised form 8 October 2013
Accepted 9 October 2013
Available online xxxx

Keywords:
Marine oil-spill detection
SAR
Classifier comparison
Bootstrap-aggregated tree-based methods
Support vector machine
Artificial neural network
Generalized additive model
Penalized linear discriminant analysis

ABSTRACT

The discrimination of oil spills and look-alike phenomena (e.g., low wind area, wind front area and natural slicks) on Synthetic Aperture Radar (SAR) images is a crucial task in marine oil spill detection. Many classification techniques can be employed for this purpose. In order to make the best use of the large variety of statistical and machine learning classification methods, it is necessary to assess their performance differences and make recommendations for classifier selection and improvement. The objective of this paper is to compare different classification techniques for oil-spill detection in RADARSAT-1 imagery. The data of this study consists of 15 features of 192 oil spills and look-alikes identified by Canadian Ice Service between 2004 and 2008 off Canada's east and west coastal areas. The studied classifiers include the Support Vector Machine (SVM), Artificial Neural Network (ANN), tree-based ensemble classifiers (bagging, bundling and boosting), Generalized Additive Model (GAM) and Penalized Linear Discriminant Analysis (PLDA). Two performance measures, the specificity at fixed sensitivity (80%) and the area under the Receiver Operating Characteristic (ROC) curve (AUC), were estimated using cross-validation to evaluate the performance of classifiers at a high sensitivity. Overall, the bundling technique which achieved a median specificity of 90.7%, at sensitivity of 80%, significantly outperformed the second best (i.e. bagging) by 1.5 percentage points, and the worst (i.e. ANN) by 15 percentage points. The median values of AUC measure indicated consistent results. Bundling and bagging achieved comparable median AUC values of about 92%, followed by GAM and PLDA, with ANN yielding the smallest. Most classifiers (SVM, bundling and especially PLDA and ANN) performed significantly better on datasets pre-processed by log-transformation and standardization than on the original dataset. These results demonstrate the importance and benefit of selecting the optimal classifiers for oil spill classification, and configuring the classifiers by proper feature construction techniques.

1. Introduction

Oil spills seriously affect marine ecosystems and cause both social and environment problems (Topouzelis, 2008). Produced by tankers or drilling platforms, marine oil spills pollute the sea water, destroy wildlife habitat and breeding ground, and damage beaches. Synthetic Aperture Radar (SAR) onboard earth observing satellites has been extensively used for oil spill detection in the marine environment due to the wide coverage and all-weather and all-day capability (Brekke & Solberg, 2005). Both oil spills and look-alike phenomena (e.g., low wind area, wind front area and natural slicks) may appear as dark formations on SAR images. It is impossible to discriminate oil spills from look-alikes solely based on SAR intensity values as oil spills assume a wide range of intensities due to their varying thickness and the complexity of the marine environment (Brekke & Solberg, 2005).

Features that further characterize the dark spots, such as geometric shape, contrast with surrounding areas and contextual information, therefore have to be extracted and used as inputs for the discrimination of oil spills and look-alikes (Brekke & Solberg, 2005; Topouzelis, 2008). Overall, there are three steps for oil-spill detection: (i) dark-spot detection to exclude most open water surfaces and identify oil-spill candidates (Li & Li, 2010; Shu, Li, Gomes, & Yousif, 2010), (ii) feature extraction for collecting ancillary features about these candidates, and (iii) classification for discriminating oil spills from look-alikes using the features extracted (Brekke & Solberg, 2005). In the final stage, it is important to achieve a high sensitivity in order to be able to respond to the vast majority of the real oil spills. The focus of this paper is therefore on oil-spill classification at high sensitivity in step (iii) of the detection process.

Several classifiers have been employed for the discrimination of oil spills and look-alikes. Solberg, Brekke, Volden, and Husøy (1999, 2007), Solberg, Storvik, Solberg, and Volden (1999), Solberg, Dokken, and Solberg, (2003) proposed a Bayesian classification scheme by combining prior knowledge, Gaussian densities and rule-based density

* Corresponding author at: Tel.: +86 592 2580003.
E-mail address: junli@xmu.edu.cn (J. Li).

0034-4257/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.rse.2013.10.012

corrections. Fiscella, Giancaspro, Nirchio, Pavese, and Trivero (2000) used linear discriminant analysis (LDA) approach based on the Mahalanobis distance. Nirchio et al. (2005) employed a multiple linear regression method for oil-spill classification. Topouzelis, Karathanassi, Pavlakis, and Rokos (2007) and Frate, Petrocchi, Lichtenegger, and Calabresi (2000) adopted the artificial neural network (ANN) approach to approximate the relation between dark-spot features and the class labels. The support vector machine (SVM) was employed by Brekke and Solberg (2008). However, these methods constitute only a limited set of popular classification techniques. Other advanced techniques such as bundling, bagging, boosting and the generalized additive model (GAM) have not been explored for oil spill classification. Moreover, a systematic, quantitative comparison of the available classifiers is still lacking, although performance differences may be substantial in their application to remote sensing problems (e.g., Brenning, 2009, Brenning, Kaden, & Itzerott, 2006; Brenning, Long, & Fieguth, 2012; Knudby, Ledrew, & Brenning, 2010).

This paper therefore aims to compare a variety of statistical and machine-learning classification techniques for oil-spill detection by using state-of-the-art evaluation methods. We employ the receiver operating characteristic (ROC) analysis (Metz, 1978; Zweig & Campbell, 1993) to evaluate the performance of the classifiers, and adopt the *k*-fold cross-validation technique for the bias-reduced estimation of performance measures. Moreover, since different classifiers require differently prepared dataset, unbiased comparison of classifiers should take into account the possible performance improvement of a classifier by suitably preparing the dataset. We therefore explore the influence of different feature preprocessing techniques on the performance of a classifier and choose the best performance for comparing with other classifiers. For the issue of feature importance which is of great interest for many practitioners, this comparison scheme allows investigation on the performance differences of a large variety of classifiers over subsets of features which tends to provide a more balanced assessment of feature importance.

2. Method

We compare 7 classifiers that predict a categorical response that includes two classes (oil spill and look-alikes) based on 15 features. We adopt the specificity at fixed sensitivity (80%) and the area under the ROC curve (AUC) to evaluate the performance of classifiers, and employ the cross-validation technique for bias-reduced estimation of performance measures. To account for the influence of data preparations on classifiers, we determine the best performance of a classifier over differently preprocessed dataset according to the AUC measure, and compare the classifiers based on their respective best performances. The significance of performance differences among classifiers is statistically tested.

2.1. Data set

In order to monitor the illegal release of oily wastes from ships traveling in Canadian waters, Canadian Ice Service (CIS) of Environment Canada has been designing a program called Integrated Satellite Tracking of Pollution (ISTOP) as part of its ice surveillance operational program towards effective use of RADARSAT images to aid oil spill detection (Gauthier, Weir, Ou, Arkett, & De Abreu, 2007). The trained human experts at CIS firstly identify the dark-spots on SAR images as oil-spill candidates. They then discriminate between oil spills and look-alikes based on their experience and prior information concerning the location, the proximity to land, the weather information, the difference in shapes, and the contrast with the surrounding sea between oil spills and look-alikes. Moreover, in order to increase the reliability of the classification, they will look at the distances between the identified oil spills and the nearest ships. Oil-spill candidates associated with ships are classified into Category 1A, which means they have the highest possibility to be true oil spills, and consequently the highest priority to be verified by aircraft. Candidates that have ships within 50 km of distance are classified into Category 1B, while those that have no ships within 50 km are classified into Category 2. Potential oil spills with relatively low confidence are put into Category 3, while those having little chance to be oil spill remain uncategorized.

The dataset of this study comprises 15 features of 192 oil spills and look-alikes identified by a human analyst at CIS based on five years (2004–2008) observations off the east and west coast of Canada (see Fig. 1). The dataset used in this study contains 93 RADARSAT-1 ScanSAR Narrow Beam images with swath width of 300 km and spatial resolution of 50 m, and covers vast Pacific and Atlantic coastal areas. Contained in these images are 98 oil spills that belong to Category 1, and 94 look-alikes (Categories 2 and 3 or uncategorized). Each image contains at least one instance of oil spill or look-like and has balanced number of oil spills and look-alikes on average. Of the 98 oil spills, 21 of them have been proved to be oil spills, but others have unknown identities due to the lack of aircraft verification. So by treating them as true oil spills to train classifiers, we are checking the ability of classifiers to approach the highest accuracy that can be achieved by human experts. Of the 94 look-alikes, 7 of them belong to Category 2, and 25 of them belong to Category 3, while the rest 62 were randomly selected from the uncategorized dark-spots that were regarded as non-oil by human analysts. For all the categorized dark-spots, their boundaries have been provided by human analysts in CIS. We therefore do not need to explicitly perform the dark-spot detection for these samples. But for the 62 uncategorized dark-spots, since no boundary information is available from CIS, we delineated their boundaries by visually discerning the gray tone difference of dark-spots and the background.

Given the dark-spots in pixel-format, features need to be extracted as input to classifiers. The features proposed by the researchers can be categorized into four groups: (i) physical and textural properties, (ii) geometric shape, (iii) contrast with background, and (iv) contextual information (Brekke & Solberg, 2005; Topouzelis, 2008). Different researchers employed different features. For example, Topouzelis et al. (2007) adopted 10 features to train neural network classifier. The number of features fall into category (i), (ii) and (iii) are respectively 5, 3, 2. Fiscella et al. (2000) used 11 features, Frate et al. (2000) used 11 and Solberg et al. (2007) used 13. In this study, we intend not to use all the proposed features by the other researchers, because it will increase the dimensionality of the dataset, thus the risk of overfitting. Therefore, we select 15 features out of all the available features as classifier input. Moreover, features that belong to the same category are highly correlated. To reduce the information redundancy, for each group of features, we select a certain number of representative features that are commonly employed by researchers in the literature (Frate et al., 2000; Solberg et al., 2007; Topouzelis, 2008). The selected 15 features describe the geometric shape (predictors no. 1–4), physical properties (no. 5–7), contrast with background (no. 8–14) and contextual information (no. 15) of identified objects, see Table 1.

(1) Target area A in number of pixels;
(2) Target perimeter P in number of pixels;
(3) Complexity measure $C = P^2/A$;
(4) Spreading measure S, the ratio between target width and length;
(5) Standard deviation of gray-scale intensity values of the object (OSd);
(6) Average intensity value of the background area (BMe);
(7) Standard deviation of the intensity value of the background (BSd);
(8) Power to mean ratio contrast (Opm/Bpm), with $Opm = BSd/BMe$, $Bpm = OSd/OMe$ and OMe representing the mean intensity value of the background;
(9) The ratio between OSd and BSd, denoted by $ConRaSd$;
(10) Local area contrast ratio $ConLa$, defined as the radio between the OMe and the mean intensity value of a window centered at the region;

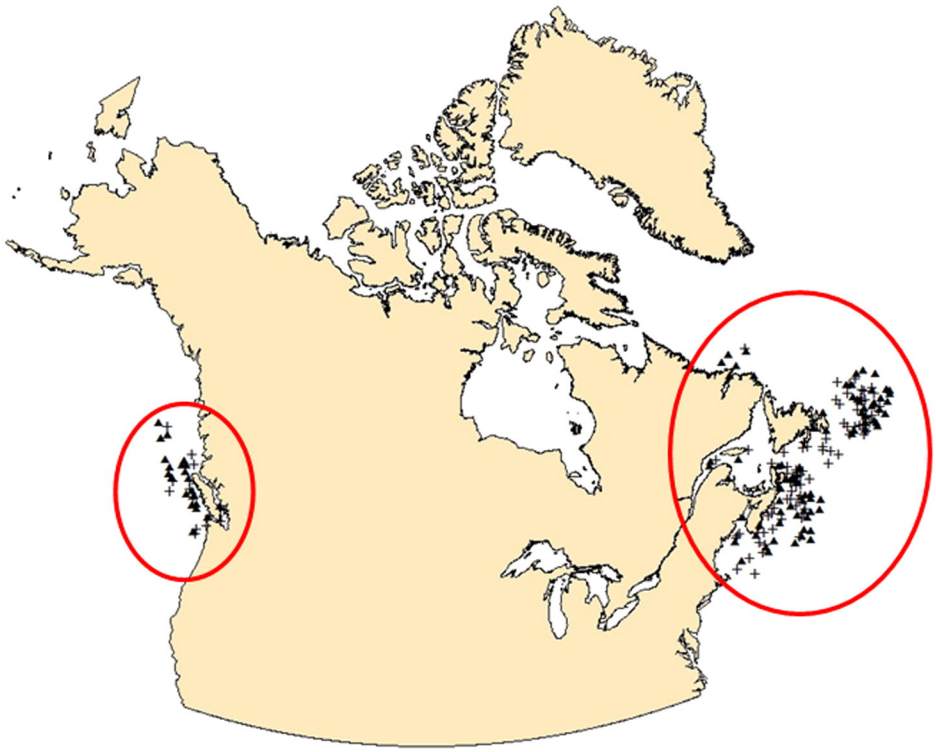

Fig. 1. The study areas (marked by the ellipses) are located off Canada's east and west coast. The identified oil spills studied in this work are denoted by symbol '+', while look-alikes are represented by symbol '▲'.

(11) Maximum contrast *ConMax*, defined as the difference between the background mean intensity value and the lowest value inside the target;
(12) Smoothness contrast: $ConSm = (N_o/G_o)/(N_b/G_b)$ where N_o is the number of target pixels, G_o: the sum of the gradient values of target pixels; N_b: the number of background window pixels; G_b: the sum of the background window gradient values;
(13) Maximum gradient value of the dark-spot border area, *GMax*. The gradients are computed by the Sobel operator;
(14) Standard deviation of the border gradient values, *GSd*;
(15) Number of neighboring targets in the same image, *N*.

2.2. Pre-processing of predictors

All the predictors used in this paper are quantitative variables. Some features (e.g., *A*, *C*, *P*, *BMe*, *N*, *ConMax*) have heavy-tailed distributions (Table 1). The predictors also have strongly varying ranges of values. Based on the characteristics of the dataset, we adopt two pre-processing techniques before training classifiers: (1) log-transform all skewed features; (2) standardize the predictors, i.e. subtract their mean value and divide them by their standard deviation. Since different classifiers prefer differently prepared datasets, "fair" comparison of classifiers should be based on the highest performances of classifiers over differently pre-processed datasets. As such the training samples are preprocessed by the different combinations of the two techniques to determine the best performance of classifiers. Afterwards we use the respective best performance of each classifier for comparison purpose.

2.3. Classifiers compared

In this study, a total of 7 selected classifiers were compared, including penalized linear discriminant analysis (PLDA), GAM, tree-based

Table 1
Summary statistics (i.e. min, median, max and interquartile range) of features for oil spills and look-alikes separately.

No.	Feature	Look-alikes				Oil spills			
		Min	Med	Max	IQR	Min	Med	Max	IQR
(1)	*A*	325	11,482	170,384	32,039	17	912	16,912	2346
(2)	*P*	106	675	2807	789	64	299	2011	289
(3)	*C*	15.2	31.5	325.9	30.6	20.4	80.0	510.8	115.2
(4)	*S*	54.9	85.5	99.9	19.7	72.1	96.4	100.0	9.7
(5)	*OSd*	7.7	19.5	46.1	10.9	8.1	20.6	52.6	17.4
(6)	*BMe*	32.2	73.0	191.0	33.9	31.6	77.3	191.7	52.0
(7)	*BSd*	11.5	28.2	56.1	16.3	10.1	28.3	56.6	21.7
(8)	*Opm/Bpm*	0.5	0.9	1.3	0.2	0.5	0.9	1.0	0.1
(9)	*ConRaSd*	1.1	1.4	2.6	0.4	1.0	1.3	2.4	0.6
(10)	*ConLa*	0.4	0.7	0.9	0.2	0.5	0.7	0.9	0.2
(11)	*ConMax*	32.2	71.0	190.9	30.0	31.6	72.9	178.6	52.2
(12)	*ConSm*	0.3	0.7	0.9	0.2	0.5	0.8	1.0	0.14
(13)	*GMax*	124.0	363.1	823.2	163.2	113.9	408.0	943.5	325.6
(14)	*GSd*	20.0	59.8	128.5	29.8	24.7	62.0	146.3	48.3
(15)	*N*	0.0	4.0	35.0	4.5	0.0	0.0	15.0	2.0

ensemble methods (bagging bundling and boosting), SVM, and ANN. Although other traditional methods (e.g., *k*-means, LDA, logistic regression and classification tree) are also possible for oil spill classification, we focus on more recently developed or adapted statistical and machine-learning techniques in this study. All analyses were performed in the R programming environment (R Development Core Team, 2005) with its contributed packages 'rpart' (Therneau & Ripley, 2010), 'ipred' (Peters & Hothorn, 2007), 'gbm' (Ridgeway, 2012), 'mda' (Hastie & Tibshirani, 2006), 'gam' (Hastie, 2006), 'LIBSVM' (Chang & Lin, 2001), 'e1071' (Dimitriadou, Hornik, Leisch, Meyer, & Weingessel, 2008) and 'nnet' (Ripley, 1996).

2.3.1. Tree-based ensemble techniques: bagging

Tree-based ensemble classifiers are computational techniques that combine a large number of individual classification trees for improved prediction (Breiman, 1996). Classification trees are very flexible classifiers that recursively split the input dataset into subsets based on binary decisions (Breiman, Friedman, Olshen, & Stone, 1984). The class label of a test object is predicted by applying the decision criteria from the root to the leaves in order to determine which leaf it falls in. Since classification trees are sensitive to the choice of particular training sample, the bagging technique has been proposed (Breiman, 1996). Instead of training the tree on the original dataset, bagging trains separate trees on many random (bootstrap) samples of the dataset. Then the class membership of a test object is decided by majority voting among the trees. The classification trees and tree-based ensemble techniques have been widely used in remote sensing classification applications (e.g., Chan & Paelinckx, 2008; Duro, Franklin, & Dubé, 2012; Miao, Heaton, Zheng, Charlet, & Liu, 2012). Bagging in particular was demonstrated by Knudby et al. (2010) the best of the six chosen classifiers for mapping of reef fish species richness, diversity and biomass. In this study, 100 bootstrap samples are used for building a committee of trees. For all trees, we use gini split criterion, 7 minimum observations in leaf node and 30 maximum depths. And we use 10-fold cross-validation with complexity parameter of 0.01 for tree pruning.

2.3.2. Tree-based ensemble techniques: bundling

Bundling is similar to bagging except that it integrates the prediction function of a classifier trained on out-of-bag samples as an additional predictor variable for building classification trees (Hothorn & Lausen, 2005). It is therefore expected to be more efficient than bagging.

In this work, 100 bootstrap samples are used for bundling. The PLDA classifier is incorporated as an ancillary classifier in bundling, using its discriminant functions as predictor variables (Brenning, 2012). In PLDA, we set the regularization parameter $\lambda = 1$. The parameters of classification tree are the same as in bagging.

2.3.3. Tree-based ensemble techniques: boosting

Boosting tree is also an ensemble technique that intends to improve the accuracy of prediction by combining the output of many tree-based classifiers. However, unlike bagging and bundling, boosting allows the evolution of trees over time and predicts the labels by weighted voting among trees. The recent uses of boosting for mapping forest biomass and global urban areas have achieved very high accuracy (Carreiras, Vasconcelos, & Lucas, 2012; Schneider, Friedl, & Potere, 2010). In this work, we adopt Friedman's gradient boosting machine approach (Friedman, 2001). We use the binomial deviance loss function. Half of the training samples are randomly selected to propose the next tree in the additive tree expansion. The shrinkage parameter is 0.01. Since the number of iterations determines primarily the generalization capability, we estimate this parameter by 5-fold internal cross-validation for efficient predictions on test samples.

2.3.4. Penalized linear discriminant analysis

The LDA predicts the class membership based on the posterior probabilities of different classes. It assumes that the densities of predictors conditioned on class membership are Gaussian and that different classes share the same covariance matrix. Then the posterior log-odds between two classes are linear function of the predictors. PLDA is designed to deal with high-dimensional data and correlated predictors by imposing smoothness constraints on the coefficients of predictors (Hastie, Buja, & Tibshirani, 1995). We use the default regularization parameter $\lambda = 1$.

2.3.5. Generalized additive model

Logistic regression, as a widely-used type of Generalized Linear Model (GLM), models the logit of class probability as a linear function of the predictors. GAM extends GLM by applying nonlinear transformation (e.g., cubic smoothing splines, or local polynomial regression) to the original predictors. Hence, GAM is more capable of modeling nonlinear correlation among variables.

In this study, stepwise forward variable selection based on the Akaike Information Criterion (AIC) is used to decide, for each predictor, whether it is omitted from the model, included as a linear predictor, or included as a nonlinear predictor that is transformed using smoothing splines of two equivalent degrees of freedom.

2.3.6. Support vector machine

The SVM nonlinearly transforms the original covariate into a higher-dimensional feature space in order to find an optimal separating hyperplane (Moguerza & Muñoz, 2006; Mountrakis, Im, & O gol e, 2011). It has been used by Brekke and Solberg (2008) for oil spill classification. Moreover, it proved to be an efficient technique in predictive geomorphological modeling (Brenning, 2005) and in land cover classification (Brenning et al., 2006; Duro et al., 2012; Foody & Mathur, 2004). *C*-classification with radial basis function is adopted in this work. The bandwidth γ of the kernel function and the regularization parameter *C* control the behavior of SVM. Instead of using the default setting implemented in the R package 'e1071', we adopt an internal 10-fold cross-validation to automatically tune the hyperparameters. Optimal hyperparameters are selected by grid search in a discretized two-dimensional parameter space along 2^d, where $d = -4, -3.5, -3, \ldots, 1$ for γ and $d = -2, -1.5, -1, \ldots, 4$ for *C*.

2.3.7. Artificial neural networks

ANNs are highly flexible tools for modeling the complex relationship between predictors and categorical responses. They provide direct estimation of the posterior probabilities of class membership (Zhang, 2000). Among many types of neural networks that can be used for classification purposes, we focus on multilayer perceptrons (MLPs), which are the most widely studied and used ANN classifiers. Because Funahashi (1998) has demonstrated that for binary *p*-dimensional Gaussian classification (here $p = 15$ features), three-layer neural networks with at least $2p$ hidden nodes can approximate the posterior probability arbitrarily well, we adopt one hidden layer and set the number of hidden nodes to be 40. The range of initial weights is set to −0.1–0.1 (Haykin, 1999; Kavzoglu & Mather, 2003). Other parameters are in accordance with the default setting in R package 'e1071': weight decay = 0; max iteration = 100; with least-squares fitting.

2.4. Accuracy measure

In this work, the analysis of ROC curves estimated by cross-validation is adopted to evaluate the performance of different classifiers. The performance of a classifier presents itself as a trade-off between true positive rate (TPR, sensitivity) and true negative rate (TNR, specificity). If a cost function is known, the optimal cut-off point that produces the smallest overall misclassification cost can be determined. Since the misclassification of true oil spills as look-alikes (expressed by the false negative rate, FNR) is more serious than the misclassification of look-alikes as oil spills (expressed by the false positive rate, FPR), it would not be appropriate to compare the classifiers based on

the misclassification error rate or overall accuracy, which assigns equal weight to FNR and FPR.

In order to do a fair comparison, the ROC analysis, which is independent of specific decision thresholds for binary prediction, is used to evaluate classifier performance. The ROC curve is a graphical plot of sensitivity and specificity as the decision threshold varies. The ROC curve of a useless classifier would follow the diagonal line, while that of a perfect classifier would follow the left and top axes of the ROC plot (Metz, 1978; Zweig & Campbell, 1993). Several techniques can reduce the ROC curve to single scalar measures, such as the area under the ROC curve (AUC), which represents the "probability that the classifier will correctly rank a randomly chosen positive instance higher than a randomly chosen negative instance" (Fawcett, 2006; Hanley & McNeil, 1982). In this paper, we use the AUC to evaluate the overall performance of classifiers.

Moreover, since it is desirable for a classifier to detect oil spills at very high accuracy, we would like to assess the ability of classifiers to correctly classify look-alikes when the accuracy of classifying oil spills is fixed at a high value, in this study 80%. This can be achieved by measuring the specificity at fixed sensitivity based on the ROC curve. The R package 'pROC' is used in this work (Robin et al., 2011).

2.5. Cross-validation estimation

To obtain unbiased accuracy estimation, the training set and test set should be independent from each other and follow the same distribution (Hand, 1997). Accuracy measures evaluated based on the training set are problematic because such measures tend to favor complex classifiers which are capable of overfitting the data, thus overestimate the ability of generalizing the learnt rule to other independent dataset. Splitting the dataset into training and test sets and estimating the accuracy measures on test set could guarantee unbiased accuracy estimation as long as the training set and test set have drawn from the same distribution. However, this approach is not suitable for limited datasets as in this study. Cross-validation can fully take advantage of the available samples by repeatedly producing training and test sets (Hand, 1997). In k-fold cross-validation (here $k = 10$), the dataset is randomly partitioned into k subsets of equal size, k-1 of which are used as a training set and the remaining subset as a test set for performance estimation. This is repeated k times so that each of the subset is used as a test set once. The performance measures are averaged over all k test sets. This procedure is repeated r times (here: $r = 100$) in order to obtain results that are independent of a particular partitioning and to be able to test the significance of observed performance differences.

When there is spatial autocorrelation among samples during cross-validation, such effects should be accounted for in order to reduce bias (Brenning, 2012). Considering the sparse distribution of oil spills over vast ocean surface in our study area, it is assumed that the observations are independent. Nevertheless, we have to assume that samples located within the same image scene are not independent because they were observed under similar environmental conditions (e.g., wind and wave regime). To be cautious, we account for this by performing the cross-validation at the image level instead of the object level. In each repetition of cross-validation, the 93 RADARSAT-1 images were randomly partitioned into 10 sets of (approximately) equal size. Since the images contain roughly equal numbers of oil spills and look-alikes, the training and test sets will also have balanced numbers of oil spills and look-alikes.

The construction of ROC curves requires numerical classifiers outputs instead of binary predictions. The classifiers are therefore set up to predict probabilities or some numeric measure of the predicted likelihood of membership in the oil-spill class. ROC curves are created for each repetition of the cross-validation procedure. The averaged ROC curve over all the repetitions for each classifier is produced by threshold averaging (Fawcett, 2006). The AUC and specificity at fixed sensitivity are extracted from the ROC curves estimated by 100-repeated 10-fold cross-validation and ranked based on their median values (Robin et al., 2011).

2.6. Statistical inference

In this work, we wish to determine whether there are significant differences between pairs of classifiers in terms of the selected performance measures. After testing the null hypothesis that the performance estimates of all classifiers are not systematically different from each other (Kruskal–Wallis test at the 5% significance level), the statistical significance of systematic pairwise differences between classifiers is determined by two-sided Wilcoxon rank-sum tests. In order to account for the problem of multiple testing, the output p-values of hypothesis tests are processed by the Benjamini-Hochberg procedure, which controls the false discovery rate (FDR) of a family of hypothesis tests (Benjamini & Hochberg, 1995). We use FDR $<= 0.05$.

2.7. Variable importance

The evaluation of the "importance" of variables is difficult due to two issues. (1) The importance of a variable may show great variation, depending on which evaluation criterion is used. As a result, features that are useless for a particular classifier may be of great help for another, while features that are useful for one classifier may become useless for another. (2) Due to the correlation effect, variables that are individually irrelevant may become relevant in the context of others, while variables that are individually relevant may be unimportant because of possible redundancies (Guyon, Gunn, Nikravesh, & Zadeh, 2006).

We adopt a recent technique called permutation-based variable accuracy importance (PVAI) to evaluate the importance of individual variables based on the degree of deterioration in the performance of a classifier if the variable is randomly permutated, or 'messed up' (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). This technique has only recently been adopted in remote sensing (Brenning et al., 2012). Comparing with univariate importance measures, PVAI considers the interaction among covariates by evaluating the importance of a variable in the context of others. To take into the account the first issue mentioned above, the PVAI technique is implemented on each of the 7 classifiers to evaluate the variable importance relative to different criteria.

We evaluate the importance of variables pre-processed by log transformation and standardization. For each partition of the 10-fold cross-validation approach, a variable is permutated 10 times and the performance deteriorations are measured by the reduction in AUC. After repeating the cross-validation 10 times, we get $10 \times 10 \times 10 = 1000$ permutations for each variable. The result is normalized by dividing it by the largest AUC reduction value. The PVAI therefore measures the relative importance of each variable (Brenning et al., 2012).

3. Results

3.1. Assessment of pre-processing methods

Employing log-transformation to pre-process the predictors with heavy-tailed distribution enabled PLDA to achieve significantly (at a FDR $< = 0.05$) better results than adopting the other three pre-processing types according to both accuracy measures (see Fig. 2). Specifically, PLDA with log-transformed predictors achieved about 5 percentage points higher specificity and AUC than with the original dataset. Measured by AUC, SVM and bundling performed significantly better when they were implemented on the log-transformed dataset.

The standardization operation alone enabled ANN to significantly outperform the case without pre-processing by about 5 percentage points in specificity and 6 percentage points in AUC. In addition, PLDA with standardized features achieved about 1 percentage point higher AUC than without pre-processing, although this is probably due to the scale-dependent property of parameter λ.

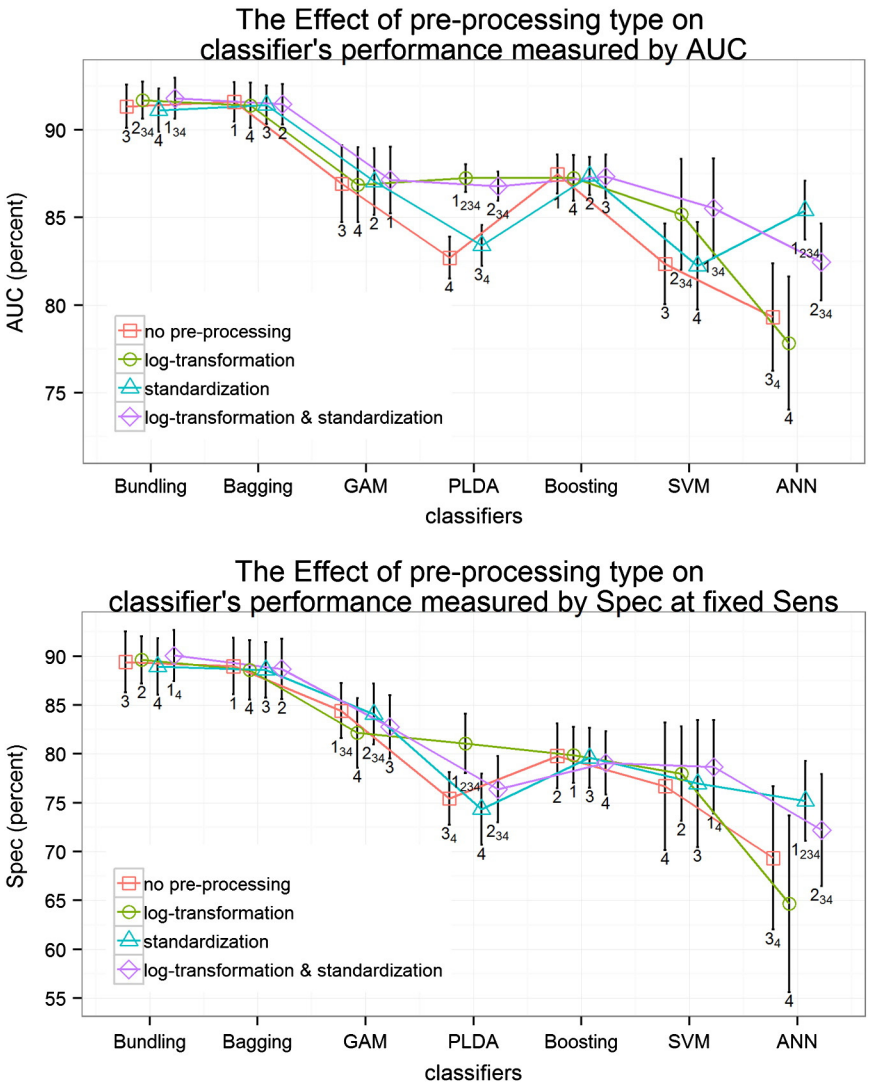

Fig. 2. Performance of the 7 classifiers on differently pre-processed dataset (no pre-processing, log-transformed, standardized, both log-transformed and standardized). The center of the bars represents the mean value, and the bars represent one standard deviation over the cross-validation repetitions. Pre-processing type with rank number x_{yzt} enabled the corresponding classifier to perform significantly (at an FDR < = 0.05) better than the pre-processing types with rank number y, z and t.

Compared with the case without pre-processing, the combined use of log-transformation and standardization to pre-process the predictors significantly increased the performances of ANN, PLDA, and SVM according to specificity at 80% sensitivity, and the performances of ANN, PLDA, SVM, and bundling according to AUC.

We identified the best performance of each classifier according to AUC measure, e.g. the best performance of PLDA is the one achieved on log-transformed dataset. And the comparison of classifiers in Section 3.2 is based on the respective best performances of classifiers.

3.2. Classifier comparison

The Kruskal–Wallis test of the null hypothesis that there are no performance differences among classifiers using the best-performing preprocessing technique was rejected at the 5% significance level (p-value: <0.001). Consequently, the two-sided rank-sum test with FDR < = 0.05 was performed on all pairs of classifiers. The results indicate that bundling achieved a specificity at fixed sensitivity that is significantly different from the other classifiers, outperforming the second best (bagging) by 1.5 percentage points, and the worst (ANN) by 14.8 percentage points (Tables 2 and 3). Both bundling and bagging achieved a median specificity of about 90%, which means that if 80% of the observed oil spills are correctly classified as oil spills, the bagging and bundling methods can still correctly classify about 90% of the actual look-alikes as look-alikes. GAM is the third best classifier, which achieved 83% median specificity. The linear method PLDA performed significantly better than the more flexible methods, i.e. boosting tree, SVM and ANN, which achieved specificities below 80%.

Bundling and bagging achieved almost identical median AUC values, followed by GAM and PLDA, with ANN yielding the smallest (see Table 2). The order obtained with median AUC is consistent with the ranking obtained with specificity, except that boosting appeared to outperform SVM.

Table 2
The median, mean and standard deviation (in %) of specificity at fixed sensitivity (80%) and AUC achieved by the 7 classifiers in 100-repeated cross-validation.

	Specificity			AUC		
Model	Med	Mean	Std.dev.	Med	Mean	Std.dev.
Bundling	90.74	90.06	2.61	91.90	91.81	1.16
Bagging	89.26	88.97	2.90	91.78	91.60	1.12
GAM	83.33	82.78	3.24	87.45	87.14	1.90
PLDA	81.48	81.06	3.02	87.33	87.25	0.79
Boosting	79.63	79.48	3.18	87.31	87.26	1.23
SVM	79.63	78.65	4.78	86.07	85.53	2.84
ANN	75.93	75.19	4.07	85.59	85.41	1.68

Table 3
Pairwise comparison of classifiers: Differences in median specificity at fixed sensitivity, tested using 21 two-sided Wilcoxon rank-sum tests. The symbol "*" indicates the difference is not statistically significant at FDR < = 0.05.

	Bundling	Bagging	GAM	PLDA	Boosting	SVM	ANN
Bundling	–	–	–	–	–	–	–
Bagging	1.48	–	–	–	–	–	–
GAM	7.41	5.74	–	–	–	–	–
PLDA	9.26	7.41	1.85	–	–	–	–
Boosting	11.11	9.26	3.70	1.85	–	–	–
SVM	11.11	9.26	3.70	1.85	0.0*	–	–
ANN	14.81	13.33	7.41	5.56	3.7	3.7	–

The performances of PLDA, boosting, bagging and bundling had relatively small variation over different cross-validation repetitions, while those of SVM, GAM and ANN had large variances, with SVM having the largest (see Fig. 3 and Table 2).

The ROC curves indicate more detailed information about the performance of classifiers (Fig. 4). No classifier could dominate the others throughout the diagram. But the bundling and bagging are generally closer to the left and top axes than the other classifiers, implying that they achieved better overall performances. More specifically, bagging performed better at low sensitivity level (0~60%), while bundling is better at high sensitivity interval (60%~100%). The other techniques show similar performances at both extremes of the ROC curve, but major differences in the middle. Particularly, boosting and ANN are closer to the left axes; PLDA is closer to the top axes, and GAM is closer to the top-left corner.

3.3. Variable importance

Some shape features (*C*, *A*) and a contextual feature (*N*) have very high PVAI values in most of the classifiers (see Table 4). Specifically, the most important feature *C* achieved the highest PVAI values in five of the seven classifiers, followed by *N* and *A* which were predominant in one classifier each. The highest PVAI values of the features related to the physical characteristics of dark-spots and the contrast of dark-spots with the background are between 0.05 and 0.61, and concentrated mostly on two classifiers, SVM and ANN. Interestingly, the composite features (e.g., *Opm/Bpm*, *ConRaSd*) did not achieve higher PVAI values than the elementary features (e.g., *OSd*, *BSd and Bme*).

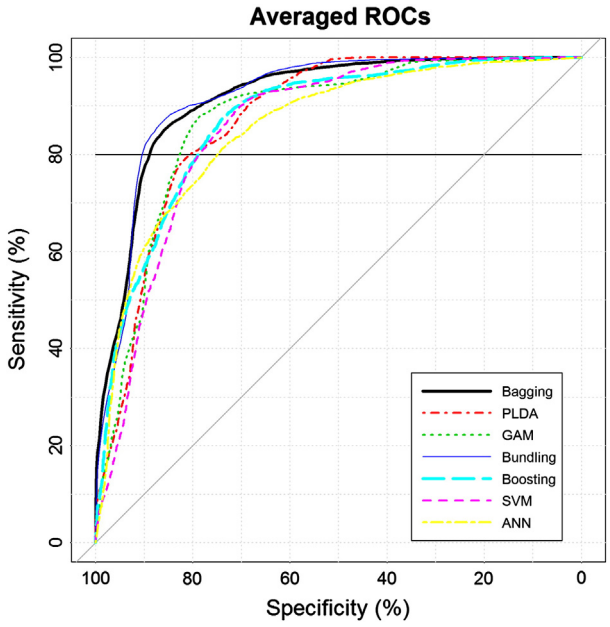

Fig. 4. The ROC curves of 7 classifiers averaged over 100 cross-validation repetitions. The horizontal black line indicates the fixed sensitivity of 80%; the diagonal gray line is the ROC curve produced by random guessing.

Different types of classifiers tended to present different patterns on feature ranking and PVAI values. The tree-based classifiers and SVM had *C*, *N* and *A* as the top three features. But the tree-based classifiers yielded very small PVAI values on the rest of the features, while SVM made relatively balanced use of all features. GAM achieved predominant usage of *A*, *P* and *N*. It is remarkable that GAM achieved zero PVAI on *C*, while all other classifiers had very high values on *C*. PLDA and ANN had *C*, *A* and *Gmax* as the top three classifiers, and differed primarily on less important features.

3.4. Label uncertainty

In this study, label uncertainty may exist due to the ambiguities in the labels of some training samples. In order to evaluate the effect of

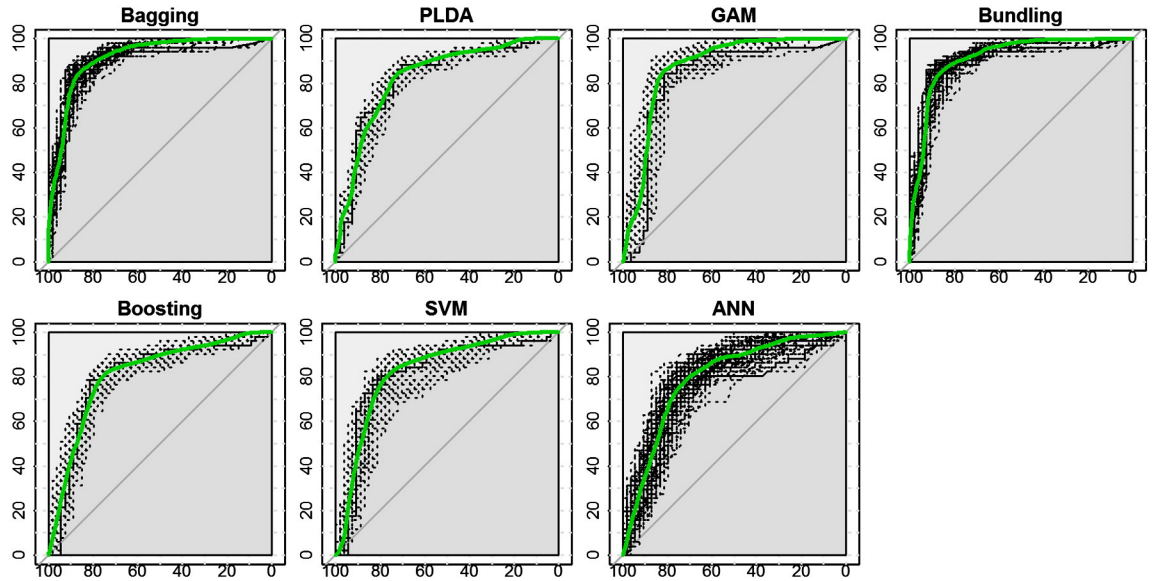

Fig. 3. The ROC curves of different classifiers. In each figure, X and Y axes are respectively specificity and sensitivity in percentage; plotted in black dotted lines are the ROC curves produced by 100 repeated-cross-validation (one line for each repetition); the green solid line is the averaged ROC curve over all the repetitions by threshold averaging; the gray diagonal line is the ROC curve produced by random guessing. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Permutation-based variable accuracy importance (PVAI) estimated by the median of AUC reductions using 10-repeated 10-fold cross-validation. Values are normalized relative to the most important predictor of each classifier. The largest value of each row is shown in bold.

	Rank	Bundling	Bagging	GAM	Boosting	SVM	PLDA	ANN
C	1	**1.00**	0.88	0.00	**1.00**	**1.00**	**1.00**	**1.00**
N	2	0.95	**1.00**	0.14	0.57	0.42	0.34	0.16
A	3	0.72	0.79	**1.00**	0.73	0.48	0.60	0.62
Gmax	4	0.08	0.00	0.00	0.03	0.27	0.59	**0.61**
BSd	5	0.03	0.00	0.00	0.04	0.21	**0.49**	0.41
S	6	0.02	0.03	0.00	0.05	**0.38**	0.01	0.09
ConMax	7	0.03	0.00	0.00	0.01	0.16	0.34	**0.35**
Consm	8	0.01	0.02	0.00	0.03	**0.35**	0.08	0.05
OSd	9	0.02	0.00	0.00	0.00	0.17	0.14	**0.27**
ConRaSd	10	0.01	0.05	0.01	0.02	**0.26**	0.04	0.01
ConLa	11	0.04	0.03	0.00	0.01	**0.21**	0.14	0.03
P	12	0.02	0.01	0.19	0.01	**0.21**	0.09	0.12
BMe	13	0.02	0.00	0.00	0.01	**0.18**	0.01	0.03
GSd	14	0.01	0.00	0.00	0.01	0.04	**0.05**	0.01
Opm/Bpm	15	0.00	0.01	0.00	0.00	0.04	0.01	**0.07**

label noise on the performance of classifiers, we plotted the conditional densities of the continuous output of the classifiers with respect to the types of training samples: (1) confirmed oil spills (i.e. the 21 verified oil spills), (2) unconfirmed oil spills (i.e. the 77 non-verified oil spills), (3) confirmed look-alikes (i.e. the 62 uncategorized dark-spots) and (4) unconfirmed look-alikes (i.e. the 32 dark-spots belong to Categories 2 and 3). We used the Gaussian kernel to estimate the density function for each of the four types of training samples identified above. We report the result of the top two classifiers which were fed by features without preprocessing. There is a large overlap between the distributions of oil spills and look-alikes in unconfirmed samples (see Fig. 5). Thus the presence of label noise may increase the difficulties for separating oil spills and look-alikes.

4. Discussion

The development of classifiers for the discrimination of oil spills and look-alikes still constitutes a big challenge. While many statistical and machine learning classification techniques can be used for this purpose, no efforts have been made to explore their suitability for oil spill detection in a comparative perspective. Our paper is the first to study the performance differences of advanced classifiers using unbiased performance evaluation techniques.

4.1. Comparison of classifiers

Overall, the bootstrap-aggregated tree-based methods (i.e. bundling and bagging) yielded significantly better results than the other methods, achieving acceptable accuracy of around 90% specificity at fixed sensitivity of 80%. We attribute this superiority to the combination of the flexibility of the tree-based techniques and the stability introduced by the bootstrap sampling. The bundling performed significantly better than bagging, indicating the improved efficiency produced by integrating an ancillary classifier (Hothorn & Lausen, 2005). While in this study they were implemented on RADARSAT-1 images, we suggest that bundling and bagging have potential to provide accurate results on dataset of other SAR sensors.

The GAM is another promising classification technique according to our study, which is theoretically capable of minimizing both the model bias by introducing nonlinear features, and model variance by selecting relevant variables in a stepwise manner. Moreover, GAM has less assumption on the distribution of predictors comparing with PLDA, and therefore more robustness to irregular distributions. In this study, most features have certain deviations from Gaussian distribution, which in addition to the possibly existence of nonlinear correlation between features and labels of dark-spots offers another explanation to the higher accuracy achieved by GAM than PLDA.

The lower accuracy achieved by boosting than other tree-based ensemble techniques, i.e. bagging and bundling is reasonable considering the particularities of boosting technique and the characteristics of our dataset. Boosting bears resemblance to bagging and bundling in that it combines the outputs of many flexible tree classifiers to produce a powerful "committee". Nevertheless, boosting is substantially different due to the fact that it allows the iterative improvement of tree classifiers and makes predictions by weighted voting among trees (Friedman, 2001). Given these particularities, boosting intends more to minimize model bias than bagging and bundling that aim primarily at reducing model variance (Carreiras et al., 2012; Ganjisaffar, Caruana, & Lopes, 2011; Maclin & Opitz, 1997). It consequently faces larger risk of overfitting during training stage, especially in the case such as our work where training samples are not abundant. This may constitute the major reason for the worse performance of boosting than bagging and bundling. Moreover, boosting is sensitive to outliers in training samples, since it gives more weights to samples that were previously misclassified. Therefore, in our work, the existence of label uncertainty in training samples may also contribute to the low accuracy of boosting.

The observation that the PLDA performed significantly better than ANN and SVM indicates that additional flexibility as provided by SVM and ANN does not necessarily improve the predictive performance compared to a less flexible linear method such as PLDA. As a "rigid" classifier, PLDA is capable of reducing model variance and providing good performance when input features have been preprocessed to roughly satisfy Gaussian distribution, as conducted in our work. On the other hand, flexible classifiers, such as ANN and SVM, trained on small-sized training samples in our work tend to overfit the dataset and could not generalize well (Atkinson & Tatnall, 1997). Since ANN assumes a large number of hyperparameters, it has proved difficult to determine experimentally the optimal values of various hyperparameters in ANN (Kanellopoulos & Wilkinson, 1997; Kavzoglu & Mather, 2003). In this work, we relied on heuristics (Funahashi, 1998; Kavzoglu & Mather, 2003) for choosing certain important hyperparameters in ANN. And

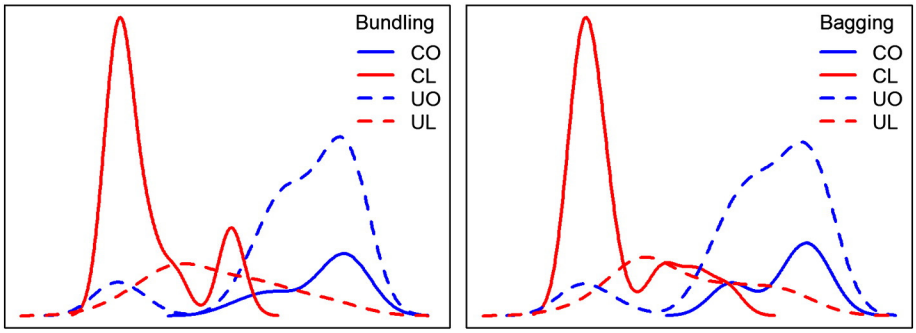

Fig. 5. The conditional densities of predicted oil-spill probabilities of two top classifiers (bundling, bagging) with respect to the types of training samples (i.e. CO: confirmed oil-spills, UO: unconfirmed oil-spills, UL: unconfirmed look-alikes, CL: confirmed look-alikes). The X-axis ranges of both plots are 0–1.

we employed internal cross-validation for tuning hyperparameters in SVM. Nevertheless, considering the small number of training samples in this study, there is no guarantee that the optimal hyperparameters values can be obtained.

4.2. Variable importance

Although most classifiers relied predominantly on limited features, they tended to present different patterns on feature ranking and PVAI values, as identified in Section 3.3. Given that the correlation effect among features has been addressed by the PVAI technique, these different patterns were primarily caused by the different preferences of classifiers. Considering the dependence of feature relevance on the characteristics of classifiers, the variation in feature importance can be better explained. For example, the higher PVAI values of *N* on tree-based classifiers than the other classifiers probably indicate that *N* is only important through complex interactions with other variables. And the zero PVAI value of *C* on GAM is probably due to some multivariate correlation that explains *C* as a linear or nonlinear function of other variables.

While the prior knowledge on features importance is of great interest to many oil-spill detection practitioners, feature importance evaluation by individual classifiers may lead to inconsistent conclusions. For example, *C*, the most "important" feature according to SVM, was the most "unimportant" to GAM. Hence, in order to reduce the bias caused by the varying preferences of individual classifiers, the identification of "important" features should be based on majority voting among many classifiers. Accordingly, we identified some geometric shape features (*C*, *A*) and the contextual feature (*N*) as predominant features, since most classifiers rely heavily on them. In contrast, only limited classifiers (i.e., SVM, PLDA and ANN) made some use of the contrast features and physical characteristics features. The oil spills in our study area are primarily attributed to tank leaking or ship washing, which may result in some typical characteristics such as small coverage and an elongated shape. In contrast, look-alikes caused mainly by low-wind areas and biogenic slicks are large and complex in shapes. That is probably why the shape features present great discriminative capability. Similarly, Topouzelis and Psyllos (2012) reported that some shape characteristics are the most important features based on the PVAI of random forest classifiers. However, using the same dataset but the ANN classifier, Stathakis, Topouzelis and Karathanassi (2006) and Topouzelis, Stathakis and Karathanassi (2009) indicated that the contrast features are the most important ones. This discrepancy that caused primarily by different classifier preferences reinforces the necessity to look at the "scores" of classifier committees for identifying the most relevant features.

4.3. Preprocessing methods

The discrimination of oil spills from look-alikes requires a data-mining system which takes into account the interaction between the pre-processing techniques and the classification models. In this study, the comparison of the performances of classifiers on dataset with different pre-processing types indicates that log-transformation can significantly improve the performance of several classifiers (SVM, bundling and especially PLDA), while data standardization can improve the performances of PLDA and especially ANN. Due to the variability of classifiers and the fact that different classifiers may require differently prepared inputs, we therefore recommend that depending on the chosen classification method, data transformations should be considered before oil-spill classification.

4.4. Accuracy measures

While misclassifying oil spills as look-alikes is more serious than misclassifying look-alikes as oil spills, most researchers applied accuracy measures without considering the different importance of FPR and FNR (e.g., Fiscella et al., 2000; Frate et al., 2000; Nirchio et al., 2005; Topouzelis et al., 2007). Since it is desirable for a classifier to detect oil spills at a high sensitivity, in this study, we evaluated a classifier by the specificity at fixed, high sensitivity level of 80%.

Since the ROC curves of different classifiers often intersect, the ranks of classifiers measured by specificity at fixed sensitivity may show variation, depending on where sensitivity is fixed. The AUC eliminates this uncertainty by summarizing the overall performance of a classifier. The combined use of specificity at fixed sensitivity and AUC therefore provides a more general assessment of classifier performances while yielding consistent results in this study.

4.5. Label uncertainty

Since the ground truth of oil spills is difficult to collect, in practice, the labels assigned by human experts are always treated as true values to train classifiers (Fiscella et al., 2000; Frate et al., 2000; Solberg et al., 1999, 2007; Topouzelis & Psyllos, 2012). The label uncertainty is a common issue in oil-spill classification considering the fact that the collection of ground truth is unavoidably affected by the subjective judgment of human practitioners. While the label uncertainty can be mitigated by more precise measurements and more careful labeling process, it can also be mitigated by employing robust models (Bouveyron & Girard, 2009; Lawrence & Schölkopf, 2001).

5. Conclusion

This paper presented a systematic comparison of popular statistical and machine-learning classification techniques for SAR oil-spill detection following the approach of Brenning (2009). According to this case study, the bootstrap-aggregated tree-based techniques bagging and bundling yielded more reliable and accurate results in oil-spill classification than all the other classifiers studied. Comparing with the worst classifier ANN, bagging and bundling methods increased the median specificity at fixed sensitivity of 80% by about 15 percentage points, demonstrating the importance and benefit of selecting the optimal classifiers for oil-spill classification. The GAM method, which introduces nonlinear features and then selects relevant features, also proved to be an efficient classifier for oil-spill identification. Boosting failed to achieve the high accuracy as by other tree-based ensemble techniques, i.e. bagging and bundling. Given the limited training dataset, a more rigid classifier such as PLDA can provide a safer alternative to flexible classifiers such as Boosting, ANN or SVM, which were prone to overfitting in oil-spill classification. For data preparation, our study demonstrated the importance of pre-processing original features by proper transformation techniques.

Acknowledgments

L. Xu was partially supported by NSF-China (#41330634 and #41374016). The authors are grateful to Dr. Ziqiang Ou of Canadian Ice Service, Ottawa, for providing the RADARSAT-1 images with identified oil spills. We are also grateful to the anonymous reviewers whose thorough feedback and suggestions have significantly contributed to improving the quality of our manuscript.

References

Atkinson, P.M., & Tatnall, A.R. L. (1997). Introduction neural networks in remote sensing. *International Journal of Remote Sensing, 18*, 699–709.
Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57*, 289–300.
Bouveyron, C., & Girard, S. (2009). Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition, 42*(11), 2649–2658.
Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140.
Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* New York: Chapman & Hall/CRC Press.

Brekke, C., & Solberg, A. H. S. (2005). Oil spill detection by satellite remote sensing. *Remote Sensing of Environment*, *1*, 1–13.

Brekke, C., & Solberg, A. H. S. (2008). Classifiers and confidence estimation for oil spill detection in ENVISAT ASAR images. *IEEE Geoscience and Remote Sensing Letters*, *1*, 65–69.

Brenning, A. (2005). Spatial prediction models for land slide hazards: Review, comparison and evaluation. *Natural Hazards and Earth System Sciences*, *5*(6), 853–862 (sRef-ID: 168 4-9981/nhess/20 05-5-853).

Brenning, A. (2009). Benchmarking classifiers to optimally integrate analysis and multispectral remote sensing in automatic rock glacier detection. *Remote Sensing of Environment*, *113*, 239–247.

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package 'sperrorest'. *2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 23–27 July 2012* (pp. 5372–5375).

Brenning, A., Kaden, K., & Itzerott, S. (2006). Comparing classifiers for crop identification based on multitemporal Landsat TM/ETM data. *Proceedings of 2nd workshop of the EARSeL special interest group on remote sensing of land use and land cover, 28–30 September, Bonn, Germany* (pp. 64—71).

Brenning, A., Long, S., & Fieguth, P. (2012). Detecting rock glacier flow structures using Gabor filters and IKONOS imagery. *Remote Sensing of Environment*, *125*, 227–237.

Carreiras, J. M. B., Vasconcelos, M. J., & Lucas, R. M. (2012). Understanding the relationship between aboveground biomass and ALOS PALSAR data in the forests of Guinea-Bissau (West Africa). *Remote Sensing of Environment*, *121*, 426–442.

Chan, J. C. W., & Paelinckx, D. (2008). Evaluation of random forest and adaboost treebased ensemble classification and spectral band selection for ecotype mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, *112*, 2999–3011.

Chang, C. C., & Lin, C. J. (2001). LIBSVM: A library for support vector machines. Software. available at. http://www.csie.ntu.edu.tw/cjlin/libsvm

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2008). *e1071: Misc functions of the Department of Statistics (e1071), TU Wien. R package version, 1, 5–18.*

Duro, D. C., Franklin, S. E., & Dubé, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*, *118*, 259–272.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Fiscella, B., Giancaspro, A., Nirchio, F., Pavese, P., & Trivero, P. (2000). Oil spill detection using marine SAR images. *International Journal of Remote Sensing*, *18*, 3561–3566.

Foody, G. M., & Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geosciences and Remote Sensing*, *42*(6), 1335–1343. http://dx.doi.org/10.1109/TGRS.20 0 4.8272 57.

Frate, F. D., Petrocchi, A., Lichtenegger, J., & Calabresi, G. (2000). Neural networks for oil spill detection using ERS-SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, *38*(5), 2282–2287.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Funahashi, K. (1998). Multilayer neural networks and Bayes decision theory. *Neural Networks*, *209–213*.

Ganjisaffar, Y., Caruana, R., & Lopes, C. V. (2011). Bagging gradient-boos ted trees for high precision, low variance ranking models. *34th Annual Association for Computer Machinery (ACM) Special Interest Group of Information Retrieval (SIGIR) Conference, Beijing, China.*

Gauthier, M. F., Weir, L., Ou, Z., Arkett, M., & De Abreu, R. (2007). Integrated satellite tracking of pollution: A new operational program. *Proceedings of the International Geoscience and Remote Sensing Symposium* (pp. 967–970).

Guyon, I. M., Gunn, S. R., Nikravesh, M., & Zadeh, L. (2006). *Feature extraction, foundations and applications*: Springer.

Hand, D. J. (1997). *Construction and assessment of classification rules.* Chichester: John Wiley & Sons.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.

Hastie, T. (2006). *GAM: Generalized additive models. R package version 0.98.*

Hastie, T., & Tibshirani, R. (2006). *mda: Mixture and flexible discriminant analysis. R package version 0.3–2, R port by F. Leisch, K. Hornik & B. D. Ripley.*

Hastie, T. J., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics*, *23*, 73–102.

Haykin, S. (1999). *Neural networks: A comprehensive foundation.* New Jersey: Prentice Hall.

Hothorn, T., & Lausen, B. (2005). Bundling classifiers by bagging trees. *Computational Statistics & Data Analysis*, *49*, 1068–1078.

Kanellopoulos, I., & Wilkinson, G. G. (1997). Strategies and best practice for neural network image classification. *International Journal of Remote Sensing*, *18*(4), 711–725.

Kavzoglu, T., & Mather, P.M. (2003). The use of backpropagating artificial networks in land cover classification. *International Journal of Remote Sensing*, *24*(23), 4907–4938.

Knudby, A., Ledrew, E., & Brenning, A. (2010). Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. *Remote Sensing of Environment*, *114*(6), 1230–1241.

Lawrence, N. D., & Schölkopf, B. (2001). Estimating a kernel Fisher discriminant in the presence of label noise. *ICML* (pp. 306–313).

Li, Y., & Li, J. (2010). Oil spill detection from SAR intensity image using a marked point process. *Remote Sensing of Environment*, *7*, 1590–1601.

Maclin, R., & Opitz, D. (1997). An empirical evaluation of bagging and boosting. *Proceedings of the Fourteenth National Conference on artificial intelligence* (pp. 546–551). Rhode Island: American Association for Artificial Intelligence Press (27–31 July 1997, Providence).

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*, 283–298.

Miao, X., Heaton, J. S., Zheng, S., Charlet, D. A., & Liu, H. (2012). Applying tree-based ensemble algorithms to the classification of ecological zones using multi-temporal multi-source remote-sensing data. *International Journal of Remote Sensing*, *33*, 1823–1849.

Moguerza, J. M., & Muñoz, A. (2006). Support vector machines with applications. *Statistical Science*, *21*(3), 322–336.

Mountrakis, G., Im, J., & O gol e, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*, 247–259.

Nirchio, F., Sorgente, M., Giancaspro, A., Biamino, W., Parisato, E., Raverar, et al. (2005). Automatic detection of oil spills from SAR images. *International Journal of Remote Sensing*, *6*, 1157–1174.

Peters, A., & Hothorn, T. (2007). *ipred: Improved Predictors. R package version 0.8–5.*

R Development Core Team (2005). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing (http://www.Rproject.org)

Ridgeway, G. (2012). Generalized boosted models: A guide to the gbm package. R package vignette. http://CRAN.R-project.org/package=gbm

Ripley, B.D. (1996). *Pattern recognition and neural networks, 1157–1174 (Cambridge).*

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., et al. (2011). pROC: An open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Schneider, A., Friedl, M.A., & Potere, D. (2010). Mapping global urban areas using MODIS 500-m data: New methods and datasets based on 'urban ecoregions'. *Remote Sensing of Environment*, *114*, 1733–1746.

Shu, Y. M., Li, J., Gomes, G., & Yousif, H. (2010). Dark spot detection from SAR intensity imagery with spatial density thresholding for oil spill monitoring. *Remote Sensing of Environment*, *19*, 2026–2035.

Solberg, A., Brekke, C., Volden, E., & Husøy, P. (2007). Oil spill detection in RADARSAT and ENVISAT SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, *45*, 746–755.

Solberg, A. H. S., Dokken, S. T., & Solberg, R. (2003). Automatic detection of oil spills in Envisat, Radarsat and ERS SAR images. *Proceedings of the International Geoscience and Remote Sensing Symposium*, *4*. (pp. 2747–2749).

Solberg, A. H. S., Storvik, G., Solberg, R., & Volden, E. (1999). Automatic detection of oil spills in ERS SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, *4*, 1916–1924.

Stathakis, D., Topouzelis, K., & Karathanassi, V. (2006). Large-scale feature selection using evolved neural networks. In Bruzzone (Ed.), *Proceedings of SPIE, image and signal processing for remote sensing XII.* (pp. 636513). http://dx.doi.org/10.1117/12.688149.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *25*. http://dx.doi.org/10.1186/1471-2105-8-25.

Therneau, T. M., & Ripley, B.A. (2010). rpart: Recursive partitioning. Available at. http://CRAN.R-project.org/package=rpart

Topouzelis, K. N. (2008). Oil spill detection by SAR images: Dark formation detection, feature extraction and classification algorithms. *Sensors*, *8*, 6642–6659.

Topouzelis, K., Karathanassi, V., Pavlakis, P., & Rokos, D. (2007). Detection and discrimination between oil spills and look-alike phenomena through neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, *4*, 264–270.

Topouzelis, K., & Psyllos, A. (2012). Oil spill feature selection and classification using decision tree forest on SAR image data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *68*, 135–143.

Topouzelis, K., Stathakis, D., & Karathanassi, V. (2009). Investigation of genetic algorithms contribution to feature selection for oil spill detection. *International Journal of Remote Sensing*, *30*(3), 611–625.

Zhang, G. P. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews*, *30*, 451–462.

Zweig, M. H., & Campbell, G. (1993). Receiver operating characteristic (ROC) plots. *Clinical Chemistry*, *39*, 561–577.