ISPRS Journal of Photogrammetry and Remote Sensing 68 (2012) 135–143

Contents lists available at SciVerse ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Oil spill feature selection and classification using decision tree forest on SAR image data

Konstantinos Topouzelis [a,*], Apostolos Psyllos [b]

[a] University of the Aegean, Department of Marine Sciences, University Hill, 81100 Mytilene, Greece
[b] European Commission Joint Research Centre, Institute for the Protection and Security of the Citizen, Italy

ARTICLE INFO

Article history:
Received 15 September 2010
Received in revised form 26 October 2011
Accepted 22 January 2012
Available online 28 February 2012

Keywords:
Oil spill
Decision forest
Feature selection
SAR
Classification
Machine learning

ABSTRACT

A novel oil spill feature selection and classification technique is presented, based on a forest of decision trees. The parameters of the two-class classification problem of oil spills and look-alikes are explored. The contribution to the final classification of the 25 most commonly used features in the scientific community was examined. The work is sought in the framework of a multi-objective problem, i.e. the minimization of the used input features and, at the same time, the maximization of the overall testing classification accuracy. Results showed that the optimum forest contains 70 trees and the three most important combinations contain 4, 6 and 9 features. The latter feature combination can be seen as the most appropriate solution of the decision forest study. Examination of the robustness of the above result showed that the proposed combination achieved higher classification accuracy than other well-known statistical separation indexes. Moreover, comparisons with previous findings converge on the classification accuracy (up to 84.5%) and to the number of selected features, but diverge on the actual features. This observation leads to the conclusion that there is not a single optimum feature combination; several sets of combinations exist which contain at least some critical features.

1. Introduction

Synthetic Aperture Radar (SAR) images are extensively used for the detection of oil spills in the marine environment, as they are independent of sun light and not affected by cloudiness. Radar backscatter values from oil spills are very similar to backscatter values from very calm sea areas and other ocean phenomena, named look-alikes (e.g. currents, eddies, upwelling or downwelling zones, fronts and rain cells). Several studies aiming at oil spill detection have been conducted (Brekke and Solberg, 2005; Del Frate et al., 2000; Fiscella et al., 2000; Karathanassi et al., 2006; Migliaccio and Trangaglia, 2004; Pavlakis et al., 2001; Stathakis et al., 2006; Topouzelis et al., 2003, 2009). A detailed introduction to oil spill detection by satellite remote sensing is given by Brekke and Solberg (2005), while a detailed comparison on the several approaches and their characteristics is given by Topouzelis (2008). Oil spill detection methodology can be summarized in four steps. First, all dark signatures present in the image are isolated. Second, features for each dark signature are extracted. Third, these features are tested against predefined values. Finally, probabilities for each candidate signature are computed to determine whether it is an oil spill, or a look-alike phenomenon.

Researchers have used different input features for oil spill classification in their studies. Several studies indicate this notice. Fiscella et al. (2000) used 14 features, Solberg and Theophilopoulos (1997) used 15 features, Solberg et al. (1999) used 11 features, many of which were different from the previous studies and in general different from the 11 features used by Del Frate et al. (2000). A general description about the calculated features is given by Espedal and Johannessen (2000), in which texture features are introduced for the first time. Moreover, Keramitsoglou et al. (2005) refer to 14 features and Karathanassi et al. (2006) use 13 features covering physical, geometrical and textural behavior. Several studies try to unify all the features used having similar characteristics (e.g. Brekke and Solberg, 2005; Migliaccio and Trangaglia, 2004; Montali et al., 2006).

The absence of a systematic research on the extracted features as well as their contribution to the classification results, forces researchers to arbitrarily select features as inputs to their systems. Previous research (Stathakis et al., 2006; Topouzelis et al., 2009) headed, for the first time, on this direction. Those studies used a combination of genetic algorithms and neural networks. The lack of the systematic research is attributed to the fact that the existing

* Corresponding author. Tel.: +30 2251036878.
E-mail address: topouzelis@marine.aegean.gr (K. Topouzelis).

0924-2716/$ - see front matter © 2012 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.isprsjprs.2012.01.005

methodologies for searching into a large number of different compilations have not been fully exploited. In this paper an effort to bridge this gap and to discover the most useful features to oil spill detection is given using decision trees forest.

A decision tree forest is a classification methodology that consists of several decision trees. Each decision tree can be seen as a decision method where its branch is taking a decision. This decision has consequences which affects its sub-branch. A decision tree (also referred to as classification, or regression tree) can be seen as a visual and analytical decision support tool, in which alternative results are calculated or a decision is taken. Trees can be "taught" to execute a command from given examples i.e. regression analysis in case the outcome is a real number or to perform a classification decision when the outcome is a class to which the data belongs. Decision trees have been widely used to remote sensing studies since the beginning of the '80s (Miller et al., 1979; Muasher and Landgrebe, 1981; Scholz et al., 1979). Lately, decision trees have been used for a variety of remote sensing subjects, like automatic land mapping (Aitkenhead and Aalders, 2011), land cover classification (AmorósLópez et al., 2011) and forest tree categorization (Yu et al., 2011).

A decision forest is an ensemble of decision trees (Fig. 1). It can be seen as one classifier which contains several classification methods or one method but various parameters of work. A new input vector is classified by each individual tree of the forest. Each tree yields a certain classification result. The decision forest chooses the classification which has the most votes over all the trees in the forest. The methodology was initially proposed by Ho (1995, 1998), Amit and Geman (1997) and later, by Breiman (2001), in an integrated form (as "random forest"). The random forest methodology contains Breiman's "bagging" idea and Ho's "random selection features". The main advantage is the estimation of the important values in the classification and the estimation of the internal unbiased error during the classification. A decision forest also estimates the relation between input variables and classification accuracy. It also computes proximities between pairs of variables, which can be used in clustering and locating outliers. Overall, decision forests mainly offer an experimental method for detecting variable interactions, and have been used in a wide variety of remote sensing applications (Baraldi et al., 2010; Clark et al., 2010; Dumas et al., 2010; Guo et al., 2011).

The present work examines the performance of a decision tree forest on a well-known problem, the oil spill detection using SAR data. The contribution to the final classification of the 25 most commonly used features in the scientific community was examined. Oil spill detection methodologies traditionally use arbitrarily selected quantitative and qualitative statistical features (e.g. area, perimeter and complexity) for classifying dark objects on SAR images to oil spills or look-alike phenomena. However, the present methodology explores the potential of selecting the most important features; thus, simplifying the classification process, yet keeping high accuracy rates. Kononenko and Hong (1997) presented some principal issues and techniques in determining which attributes (features) are important for modeling and classification. They showed that classification accuracy can be improved by computing quality measurements from the available solutions.

The paper is organized in six sections. After the present introduction, a theoretical description of the decision forest is given, followed by a detailed description of the used dataset in Section 3. In Section 4 results are presented. The evaluation of the decision forest contribution is given in Section 5 and in last section, results are discussed and some conclusions are drawn.

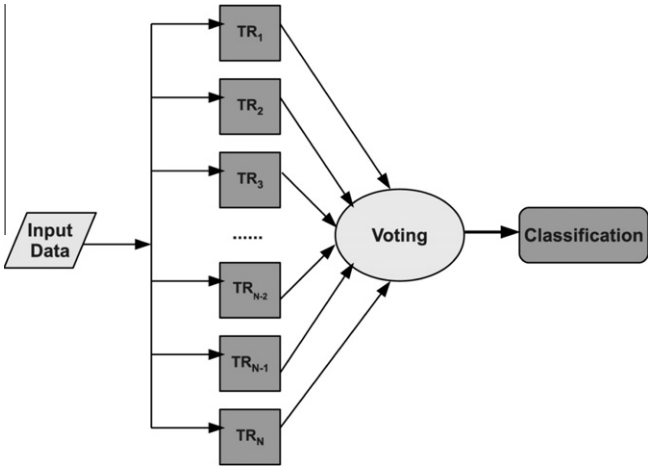

Fig. 1. Principle of decision tree classification using N decision trees (TR).

2. Decision trees and bagging

A decision forest can be seen as a group of decision trees. The latter are classification tools that use a tree-like graph structure. The feature vector is split into unique regions, corresponding to the classes, in a sequential manner (Breiman et al., 1984). Presenting a feature vector, the region to which the feature vector will be assigned, is searched via a sequence of decisions along a path of nodes of an appropriately constructed tree. The sequence of decisions is applied to the individual features and the questions to be answered are of the form $X > C_j$ where C_j is a proper threshold value or for categorical queries, when $X \subset A$.

Such trees are known as ordinary binary classification trees (OBCT). Given an input feature vector $X, X \in R^n$, a binary decision tree is built with the following steps.

2.1. Binary questions

A set of binary (true/false) questions are asked, of the form: $X \subset A, A \subseteq X$, or $X > C_j$. For each feature, every possible value of the threshold C_j defines a specific split of the subset X. In theory, an infinite set of questions has to be asked; but in practice, only a finite set of questions can be considered leading to the best split of the associated subset. The best split is decided according to a splitting criterion.

2.2. Splitting criterion

Every binary split of a node generates two descendant nodes. A criterion for tree splitting t is based on a node impurity function $I(t)$. A variety of node impurity measures is defined, as shown in Eq. (1).

$$I(t) = \varphi(P(\omega_1|t), P(\omega_2|t), \ldots, P(\omega_M|t)) \quad (1)$$

where φ is an arbitrary function and $P(\omega_i|t)$ denotes the probability that a vector X_t belongs to the class $\omega_i, i = 1, 2, \ldots, M$. A usual choice for φ is the entropy function from Shannon's Information Theory, as shown in Eq. (2).

$$I(t) = -\sum_{i=1}^{M} P(\omega_i|t)\log_2 P(\omega_i|t) \quad (2)$$

where $\log_2$ is the logarithm with base 2 and M is the total number of classes. The decrease in node impurity is defined as shown in Eq. (3).

$$\Delta I(t) = I(t) - a_R I(t_R) - a_L I(t_L) \quad (3)$$

with a_R, a_L the proportions of the samples in node t, assigned to the right node t_R and the left node t_L ,respectively. The task now reduces

to one of adopting, from the set of candidate questions, the one that performs the split, leading to the highest decrease of impurity according to Eq. (3). It should be noted that the properties of the resulting final tree seem to be rather insensitive to the choice of the splitting criterion (Breiman et al., 1984).

2.3. Stop-splitting rule

A simple stop-splitting rule has been adopted: when the maximum value of $\Delta I(t)$, over all possible splits, is less than a threshold T then splitting is stopped. Other alternatives are to stop splitting either when the cardinality of the subset X_t is small enough or when X_t is pure, in the sense that all points in it belong to a single class (Theodoridis and Koutroumbas, 2006).

2.4. Class assignment rule

Once splitting is stopped, a node is declared to be a leaf, and a class label ω_j is given using the majority rule:

$$j = argmaxP(\omega_i|t) \quad (4)$$

In other words, a leaf t of the tree is assigned to the class where the majority of the vectors X_t belong to. In our problem two classes exist: oil spill or look-alike. Therefore, each leaf votes to one of the two classes on the dependant of the majority of the vectors $\backslash X_t$.

A critical factor in designing a decision tree is its size: it must be large enough, but not too large; otherwise it tends to learn the particular details of the training set and exhibits poor generalization performance. Experience has shown that the use of a threshold value, for the impurity decrease as stop-splitting rule, does not always lead to optimum tree size. Many times, it stops the tree growing either too early or too late. The most commonly used approach is to grow a tree up to a large size first and then prune its nodes according to a pruning criterion (Mingers, 1989). Tree size has a significant importance to the present study since it is dealing with a two-class problem. Trees too large or too small will incorrectly represent the feature vectors.

A drawback associated with tree classifiers is their high variance. In practice it is not uncommon for a small change in the training data set to result in a very different tree. The reason for this lies in the hierarchical nature of the tree classifiers. An error that occurs in a node close to the root of the tree propagates all the way to the leaves. This problem is more essential for a two-class problem. Discrimination between oil spills and look alike requires careful development of training sets since an error in the tree's root will classify wrongly all the feature vectors. In order to make tree classification more stable a technique called "Bagging" has been invented (Breiman, 2001).

Bagging, which stands for "bootstrap aggregation", is a type of ensemble learning introduced by Breiman (1994), in order to improve the accuracy of a weak classifier by creating a set of classifiers. In this method, each classifier's training set is generated by randomly drawing N examples, with replacement, with N the size of the original training set. The learning system generates a classifier from the sample and aggregates all the classifiers generated from the different trial to form the final classifier.

To classify an instance, every classifier records a vote for the class to which it belongs and the instance is labeled as a member of the class with the most votes. In case that more than one class jointly receives the maximum number of votes, then the winner is selected at random. Every tree in the ensemble is grown on an independently drawn bootstrap replica of input data. Observations not included in this replica are "out-of-bag" for this tree (Breiman, 1994).

The prediction error of the bagged ensemble is estimated by computing predictions for each tree on its out-of-bag observations; averaging these predictions over the entire ensemble for each observation and then comparing the predicted out-of-bag response with the true value at this observation. Bagging works by reducing variance of an unbiased base learner, such as a decision tree. This technique tends to improve the predictive power of the ensemble, as the random selection of features reduces the correlation between trees in the ensemble.

2.5. Outliers

In order to verify further the classification rate, an outlier measure is defined. Outliers are generally defined as cases that are removed from the main body of the data. In the present study, an outlier measure helps to understand how feature combinations treat measurements, i.e. it examines the proximity of the decision forest results with the ground truth data. An outlier is calculated as the inverse of the average squared proximity between an observation and the rest of the data. Then, it is normalised by subtracting the median of the distribution; the absolute value of the result is finally divided by the median absolute deviation (Breiman, 2001).

3. Dataset for oil spill detection

The proposed methodology has been evaluated an image dataset which derived from 24 high resolution SAR images from the European Remote Sensing Satellite 2 (ERS-2). The dataset contains PRI SAR products with 100 km swath width and spatial resolution of 25 m × 25 m. It includes several sea states e.g. calm sea, small gravity-capillary waves and long swell waves. All images contain several dark objects and have been previously used by different researches (Karathanassi et al., 2006; Pavlakis et al., 2001; Stathakis et al., 2006 and Topouzelis et al., 2009). To speed up the process, the method is applied on selected windows rather than complete scenes.

From the 24 SAR images, 159 image windows are extracted, containing 90 look-alikes and 69 oil spills. For each dark object selected in the above mentioned image windows, a set of 25 features is extracted. Note that the operation is carried out on objects rather than pixels. Dark formations are detected using an object-oriented methodology described in detail in Topouzelis et al. (2003) and Karathanassi et al. (2006). Dark formations were classified manually to oil spills or look-alikes by visual interpretation. The size of the extracted windows was not constant. It varied according to the size of the dark formations. Three skilled operators analyzed the reference data and only those in agreement were used for further analysis. The operators' ability to distinguish oil spills from lookalikes was based on their experience which in later stages were transformed to a mathematical–numerical measurement of features (e.g. size: can not be very big or too long, complexity: not very complex and usually with a narrow shape, backscatter value: significant difference with the neighbouring area, etc.).

Features can be generally grouped in three major categories (Karathanassi et al., 2006; Solberg et al., 2007; Topouzelis et al., 2003; Topouzelis, 2008); features referring to the geometrical characteristics of oil spills (e.g. area, perimeter, complexity), features capturing the physical behavior of oil spills (e.g. mean or max backscatter value, standard deviation of the dark formation or a bigger surround area) and features referring to the textural behavior of the oil spills (e.g. spectral texture).

Table 1 presents a grouping of the 25 most commonly used features applied in the majority of research studies as well as in the present study. The first six features (1–6) refer to the geometrical characteristics, the next 16 features (7–22) refer to the physical characteristics and the last three (23–25) to the texture

Table 1
Commonly used features (after adaption from Stathakis et al. 2006).

No	type	Features	Code
1	Geometrical	Area	A
2		Perimeter	P
3		Perimeter to area ratio	P/A
4		Complexity	C
5		Shape factor I	SP1
6		Shape factor II	SP2
7	Physical	Object mean value	OMe
8		Object standard deviation	OSd
9		Object power to mean ratio	Opm
10		Background mean value	BMe
11		Background standard deviation	BSd
12		Background power to mean ratio	Bpm
13		Ratio of the power to mean ratios	Opm/Bpm
14		Mean contrast	ConMe
15		Max contrast	ConMax
16		Mean contrast ratio	ConRaMe
17		Standard deviationcontrast ratio	ConRaSd
18		Local area contrast ratio	ConLa
19		Mean border gradient	GMe
20		Standard deviation border gradient	GSd
21		Max border gradient	GMax
22		Mean Difference to Neighbors	NDm
23	Textural	Spectral texture	TSp
24		Shape texture	TSh
25		Mean Haralick texture	THm

characteristics of the dark formations. Detailed descriptions can be found at Stathakis et al. (2006) and Topouzelis et al. (2009).

The dataset consists of 159 samples with classified oil spills (labeled '0') and look-alikes (labeled '1'). These samples are derived from the photo interpretation result, namely, ground truth labels which evaluate the proposed recognition scheme. The dataset is divided randomly in two equal parts, one for training and the other for classification purposes. As the purpose of the paper is to identify the combination of features which perform better during classification, the term Principal Feature (PF) can be used to define those features participating in the optimum solution.

Notice that previous experiments (Karathanassi et al., 2006; Topouzelis et al., 2003) tried to rank features according to their importance. The objective was to select and use only those features characterized by strong discriminative capacity. It appears that combinations of features with high discriminative capacity were not giving as satisfying results as combinations of features having lower discrimination capabilities. For example, if we have 10 cases, three of which can be discriminated by feature X, the question is how many of the remaining seven cases can be discriminated by another feature e.g. feature Y. If features X and Y contribute only in discriminating the same three cases, then the combination is not good and another feature (e.g. feature Z) has to be used.

4. Results

A variable forest sizehas been tested including up to 300 trees, thus formatting a decision forest. Fig. 2 gives the performance of classification, in terms of classification error. Notice that for a forest containing 70 trees, the classification error becomes minimum, approximately 18%. This result reflects the fact, that above one limit, adding new trees to the forest, does not contribute significantly to the classification rate; instead, it add fuzziness in the voting process. In the following classification steps, a forest size of 70 was adopted forminimising error and to speed-up calculations.

The most reliable measure of feature importance is based on the decrease of classification accuracy when the values of a variable in a node of a tree are permuted randomly with the out-of-bag ones (Breiman et al. 1984). For every tree grown in the forest, the number of votes cast for the correct class classified was counted and the values were randomly permuted with those left out-of-bag. The average of this number, over all the trees in the forest, formulates the importance score for every feature. Finally, this measure is divided by the standard deviation over the entire forest.

In Fig. 3, the feature importance is plotted against the names of the features employed. Features are grouped according to their nature i.e. geometrical, physical, textural and ordered according to their importance. Also, the three most important combinations with 4, 6 and 9 PFs are drawn with thick black lines. As it can be seen, four geometrical features present the higher importance. The PFs formatting the 4 PFs solution are C, P, SP2, SP1. Their importance can be separated in two pairs: those (C and P) with having importance values around 0.65 and those (SP2, SP1) having importance values above 0.55, but less than 0.60. The following solution, the 6 PFs solution, contains another two features (ConLa and Bpm), whose importance values are more than 0.45 but less

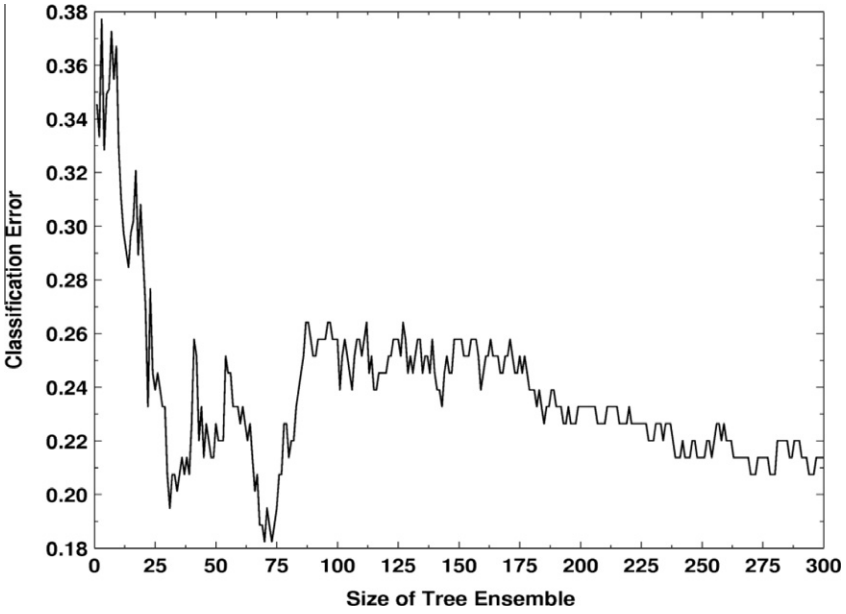

Fig. 2. Error of classification based on 25 features against forest size.

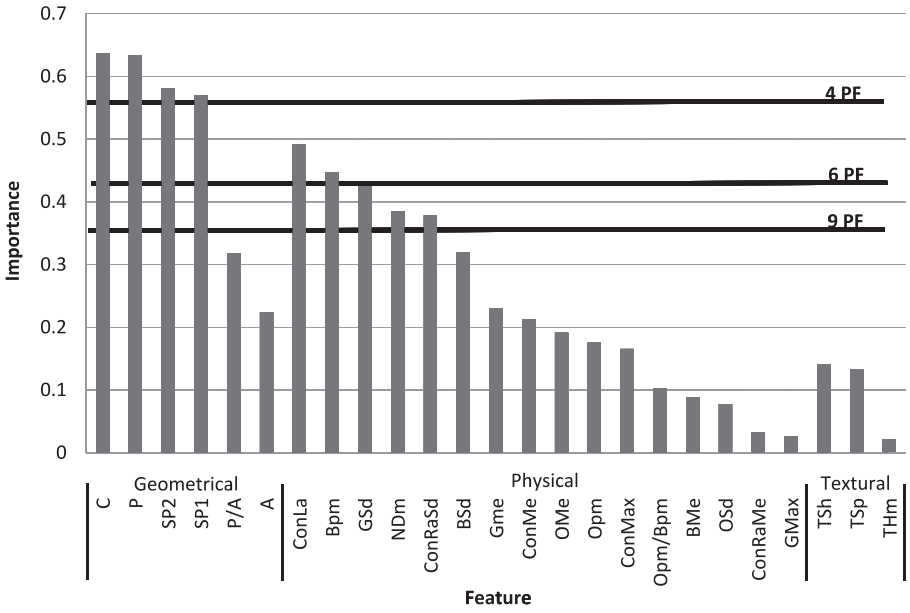

Fig. 3. Relative feature-importance for oil spill classification parameters. Horizontal black lines denote important combinations; their numbers indicate the number of Principal Features (PFs) in the combination.

than 0.50. Therefore, the importance value of 0.45 indicates the threshold value among 4 and 6 PFs solutions. The solution of 9 PFs includes three more features, theGSd, Ndm, and ConRaSd with importance values higher than 0.35. Another solution could be derived from those PFs having an importance value higher than 0.3. That solution includes eleven PFs. Those are the previous nine features and the two new onesP/A and BSd. From the 25 features under examination, eight (A, Gme, ConMe, OMe, Opm, ConMax, TSh and TSp) have similar and not significant importance values between 0.10 and 0.3. The remaining six features (Opm/Bpm, BMe, OSd, ConRaMe, Gmax, and THm) are the least important, with values less than 0.10. These importance values concern separation of the dark formations in portion of SAR images into two classes: oil spills and look-alikes.

An outlier measure calculated in order to verify the classification rate. A high value of an outlier measure indicates that this observation has a large difference between the forest result and the ground truth measurement. PFs combinations that present outliers with high values are considered worse than those having outliers with small values. Fig. 4 plots the outlier measure against the respective number of ground truth observations for the three dominant PFs combinations; i.e. those having 4, 6 and 9 PFs and the solution with all features together, i.e. 25 PFs. The 4 PFs combination presents the larger number of cases with 0 outlier value, but also the larger number of outliers with high values. Moreover, it is the only combination having outliers with a value of 10. The combination of 25 PFs presents a significant number of outliers with values of 2, while only the 6 and 9 combinations have most of their outliers with values 0 and 1. The latter two combinations have very similar behavior and none can obtain significant advantage due to outlier measurement.

Fig. 5 plots the classification and class separation for the two principal data set coordinates, i.e. the two principal components of each solution. This figure contains the three dominant combinationsand the solution with all features together. Note that the 4 PFs solution is the worst because the two categories oil spill and look-alikes are mixed in a large extend. The only part of the data which can be easily separated is a small portion in the northeast part of the 4 PFs graph. The 25PFs solution exhibits the opposite behavior. Separation ability has been significantly improved, a small portion of data is mixed and the two classes are well separated. The last two observations enhance the theory indicating that 4 features are not enough to separate correctly the given database and 25 are large enough to generalize with low accuracy the validation set, since the features have been over trained during the training phase. The remaining two solutions, i.e. those of 6 and 9 PFs, have similar behavior and arethe solutions with the highest separation ability. They both have portions covered totally with one class and the majority of the database can be separated. Nevertheless, in the central part of the graphs several cases exist with both classes together, not permitting the decision to a combination with higher separation ability. Note that the 9 PFs combination have slightly better separation ability than the 6 PFs combination since the two classes are spread wider into the graph. Therefore, the 9 PFs combination has the most appropriate behavior in separating the two classes.

Fig. 6 plots a Receiver Operating Characteristics (ROC) curve for 4, 6, 9 and 25 PFs combinations. The ROC curve for a binary classifier system is defined as the fraction of true positive rate versus the false positive rate, when the discrimination threshold is varied. Combinations with better performance should be presented in the upper left part of the diagram. Note that solutions with 6 and 9 PFs have much better performance than those with 4 and 25 PFs. This result is attributed to the fact that decreasing the PFs combinations leads to an over-simplification of the decision forest. On the other hand, increasing the PFs could possibly make the decisions very depended on a particular observation and thus non-generalising. Among the two well performed combinations of 6 and 9 PFs no significant observation can be made. Both are presenting similar behavior and have almost identical values in the critical cross point with the hypothetical diagonal line among the higher values of the axes.

Table 2 presents the classification accuracy for 4, 6, 9 and 25 PFs respectively. Note that the 6 PFs solution presents the maximum classification accuracy of 85%, which is followed by the 9 PFs solution with 84.4% accuracy. The remaining two combinations present significantly lower classification performance. The three first combinations present higher performance than all features together i.e. 25 PFs solution. This can be explained by the Hughes effect (Kavzoglou and Mather 1999), which states that for a given

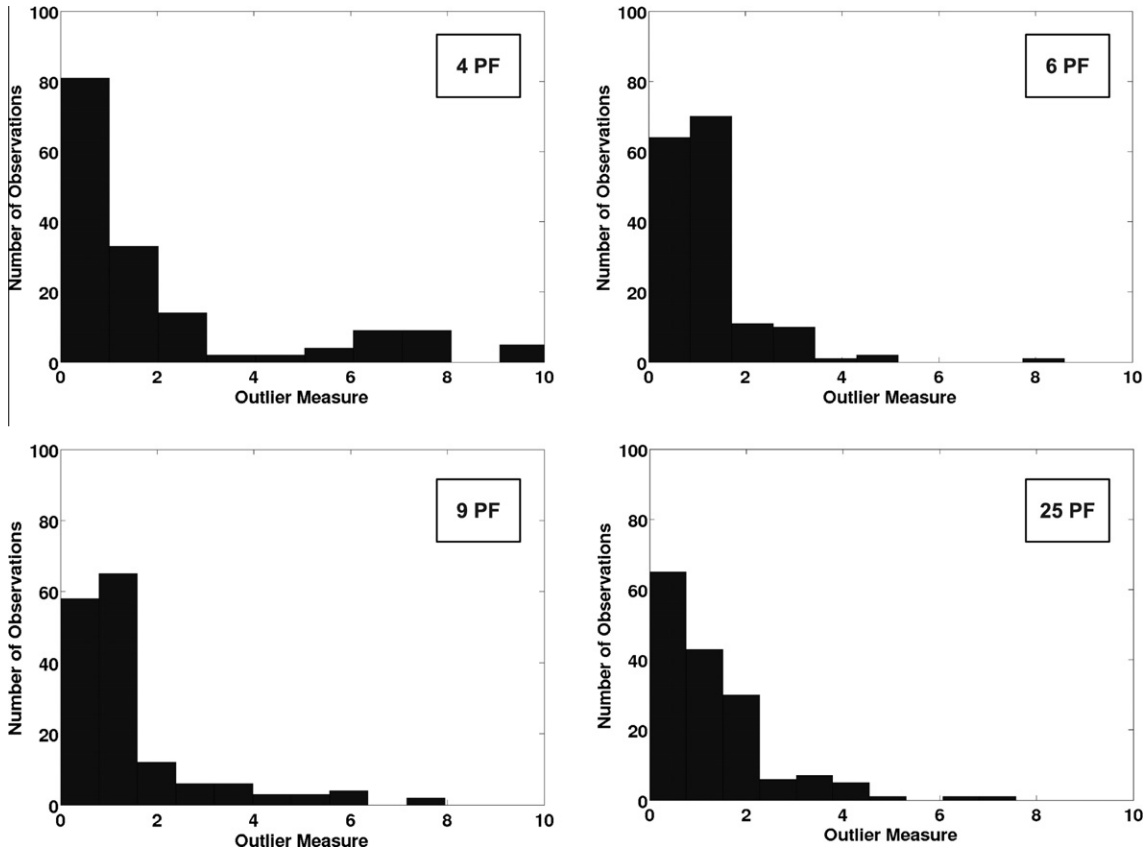

Fig. 4. Frequency histogram for outlier measure relative to the number of observations of the three important Principal Features (PF) combinations, complemented by the combination of all features.

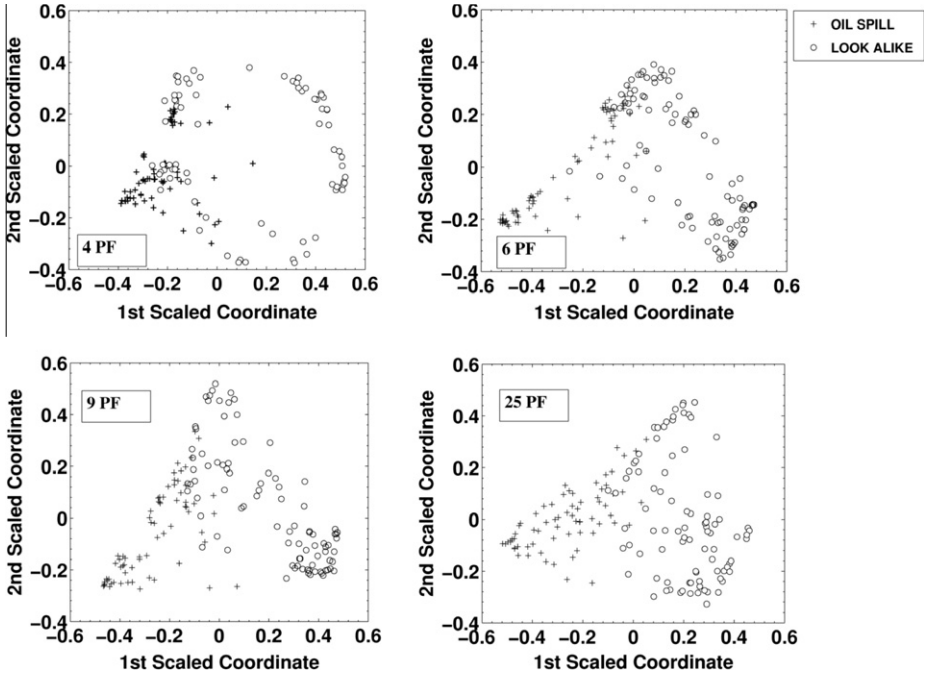

Fig. 5. Classification and class separation for the two principal data set coordinates.

number of training samples, a larger network may have a poorer generalization capability than one with fewer inputs.Hence, the inclusion of multiple features in the network input often results in the network being effective in classifying the training data, but being less effective in identifying cases that have characteristics that differ from any of the training samples. Conversely, a small

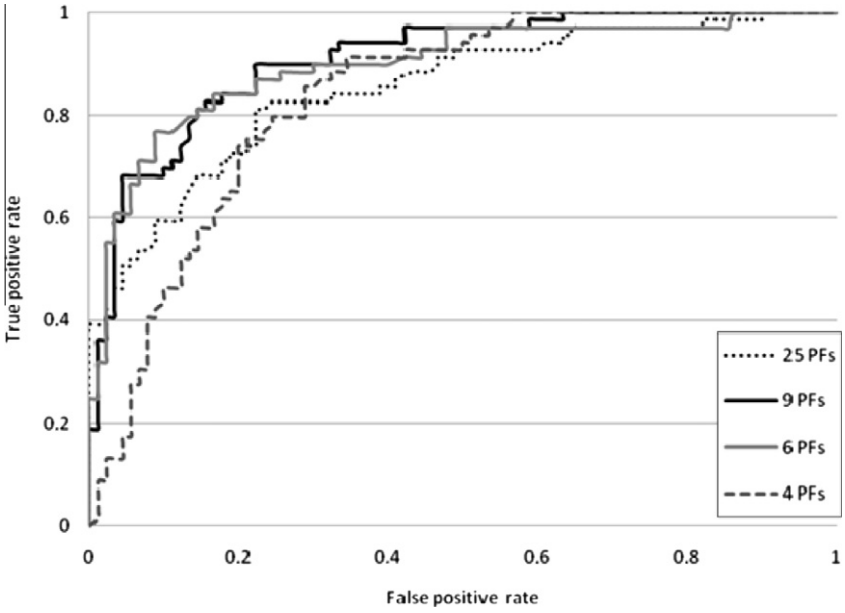

Fig. 6. Receiver Operating Characteristic (ROC) curve for the solutions under examination.

Table 2
Principal features selected from decision forests.

Number of features	Principal features (Sorted in decreasing importance)	Classification accuracy (%)
4	C, P, SP2, SP1	77.4
6	C, P, SP2, SP1, ConLA, Bpm	85.0
9	SP2, P, SP1, C, ConLA, Bpm, GSd, Ndm, ConRaSd	84.4
25	SP2, P, SP1, C, ConLA, Bpm, GSd, Ndm, ConRaSd, P/A, Bsd, Gme, A, ConMe, Ome, Opm, ConMax,Tsh, Tsp, Opm/Bpm, Bme, OSd, ConRaMe, Gmax, Thm	79.3

Table 3
Methods comparison of the feature solution.

Method	Feature solution	Accuracy(%)
Decision Forest	SP2, P, Bpm, ConLa, GSd, SP1,NDm, ConRaSd, C	84.4
Decision Forest	All features	79.3
Computational intelligence	SP2, BSd, ConLa, Opm, Opm/Bpm, OSd, P/A, C, Thm	84.8
Computational intelligence	All features	77.2
SFFS (Fisher)	A, SP2, P, ConLa, ConMax, GMe, NDm, OMe, C	83.5
SFFS (Divergence)	A, SP2, P, ConLa, ConMax, GMe, NDm, OMe, C	78.5
SFFS (Euclidean)	SP2, Bpm, BSd, ConLa, ConRaSd, Gmax, GSd, NDm, Opm/Bpm	69.6
SFFS (Mahalanobis)	A, SP2, ConLa, ConMax, GMe, NDm, OMe, C, Thm	82.3
SFFS (Bhattacharyya	A, SP2, P, ConLa, ConRaMe, SP1, Opm/Bpm, P/A, C	81.0
SFFS (Patrick Fischer)	A, P, ConLa, ConMe, ConRaMe, GMe, NDm, OMe, Thm	74.7

network may generalize well, but from an incomplete base, since the locations of decision boundaries may not be properly determined from a small number of features. Therefore, the 4 and the 6 PFs solutions can be questioned in larger data set because the locations of decision boundaries may not be properly determined. Moreover, in previous studies (Topouzelis et al. 2009) a combination with 6 features was never reported to be more effective than those having 9–10 features. Having that in mind, and taking into account that the 9 PFs solution presents very similar behaviour to the 6 PFs solution, the 9 PFs solution can be seen as the optimum feature combination of the decision forest examination.

5. Evaluation of the decision forest contribution

The above results are compared with those of previous studies (Stathakis et al. 2006, Topouzelis et al. 2009) of similar oil spills and look-alikes datasets. Those studies included computational intelligence study, referring to the synergetic use of neural networks and genetic algorithms. The genetic algorithm identified the input feature combination that gives the highest classification accuracy; while neural networks were used as a classifier. One of the best performing solutions included 9 features, shown in Table 3.

Moreover, in order to examine the robustness of the proposed feature combination, a comparison with the results of several

commonly used separability indices, including Euclidian, Fisher and Mahalanobisdistance was previously performed (Topouzelis et al. 2009). The Sequential Forward Floating Selection (SFFS) algorithm (Pudil et al. 1994) was used for comparison reasons. SFFS proceeds by successively including and excluding a variable (floating) number of features approximating the optimal solution as much as possible.

Table 3 contains the nine feature solutions of the decision forest, the computational intelligence and the several commonlyused separability indices that are deployed to guide the sequential floating forward selection. A close look to the nature of the selected features, reveals a convocation on their characteristics. Regarding the9 PFs solution of the decision forest with accuracy of 84.4%, four of the selected features referred to geometrical characteristics (SP2, P, SP1, C), five to physical characteristics (Bpm, ConLa, GSd, NDm, ConRaSd) and none to texture characteristics. The computational intelligence study (Stathakis et al. 2006 and Topouzelis et al. 2009) resulted in a combination of nine features with accuracy 84.8%, in which three were from geometrical characteristics (SP2, P/A, C), five from physical characteristics (BSd, ConLa, Opm, Opm/Bpm, OSb) and one from textural characteristics (THm). Furthermore, the SFFS best performance solution with accuracy of 83.5% is given by the Fisher criterion, in which four features referred to geometrical characteristics (A, SP2, P, C) and five to physical characteristics (ConLa, ConMax, GMe, NDm, OMe). Comparing the results of the three above mentioned solutions, three features are common i.e. C, SP2 and ConLa. This observation is encouraging for the robustness of those features.

Nevertheless, SFFS indexes present low to medium correlation with decision forest results. The Euclidean criterion is the only one having six common features with the decision forest result but it presents 14% less classification performance. Among the remaining five SFFS indexes, three have five common features with the decision forest solution, one i.e. Mahalanobis has four common features and one i.e. Patrick Fischerhas three common features. Also, the performance of using all the 25 features is 79.3% for decision forest and 77.2% in computational intelligence results.

6. Conclusions and discussion

This paper uses a very efficient classification method, a decision forest, in order to evaluate important features which discriminate oil spills from look-alikes. The separation ability of 25 features was examined and a combination of 9 features was found to be very efficient with overall classification accuracy of 84.4%. The results are quite encouraging because the proposed feature combination achieves higher classification accuracy than standard sequential selection methods.

It is worth mentioning that the classification accuracy rises when a portion of 25 features is used. This finding conforms to previous observations that feature selection can result in accuracy improvement due to Hughes effect (Kavzoglu and Mather, 1999). This effect explains the poorer generalization capability of a classifier with large number of inputs against the same classifier with smaller number of inputs.

The proposed combination of nine features was compared with previous findings using computational intelligence i.e. synergetic use of neural networks and genetic algorithms (Stathakis et al., 2006). That study resulted in another combination of nine features with overall accuracy of 84.8%. Both accuracies are very similar and the 0.4% difference could occur due to several fragile factors, e.g. the stop learning criterion in decision forest, the architecture decision in neural networks or the random selection of training and testing samples. In addition, both studies achieved higher accuracy than well-known statistical indexes. We easily conclude that both methodologies converge on the number of the used features. In both studies, 25 features were examined and nine of them were found to present very high classification accuracy. Nevertheless, in both studies some paradox observations were made. In decision forest, a six feature combination gave the highest accuracy of 85.0%; while in computational intelligence, a 10 feature combination gave 99.5% accuracy. These observations can be seen as outliers (or not stable solutions) as e.g. for neural networks that accuracy was never reproduced when a new random separation made to the used database on training and validation sets. For the six features solution of the decision forest, its generalization ability is questioned in larger data set since the locations of decision boundaries may not be properly determined due to the small number of features.

On high importance is the same number of features, i.e. nine features, resulted from both studies. While the number of features converge to nine, only three of them are common to both studies. Among the three common features, two are coming from geometrical features i.e. complexity (C) and shape factor 2 (SP2) and one from physical characteristics i.e. local area contrast ratio (ConLA). While those features seem to be crucial for oil spill discrimination, questions arise concerning the use of the remaining six features. A possible explanation is that these three features discriminate the vast majority of the cases and the rest are adapted to the less important features. This probably happens because the discrimination capability can be explicitly measured for one feature only. It is well understood that in the multiple feature case the total discrimination capability is not equal to the sum of the individual features. On the contrary, the total discrimination capacity of a combination is consummative to the discrimination capabilities of each feature. Therefore, it can be concluded that there is not a single combination which can give extremely high results, but probably a set of combinations exist which result in the same high level. These sets contain critical features, which discriminate most of the cases. In our study those features are the three common features.

Although it is not easy to understand the effect of combining features, another conclusion holds. The separation of features to three categories i.e. geometrical, physical and textural, is a significant step for further research. Further studies could include more features for evaluation as well as a larger data set of verified oil spills and look-alikes.

References

Aitkenhead, M.J., Aalders, I.H., 2011. Automating land cover mapping of Scotland using expert system and knowledge integration methods. Remote Sensing of Environment 115 (5), 1285–1295.

Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural Computation 9 (7), 1545–1588.

Amorós López, J., Izquierdo Verdiguier, E., Gómez Chova, L., Muñoz Marí, J., Rodríguez Barreiro, J.Z., Camps Valls, G., Calpe Maravilla, J., 2011. Land cover classification of VHR airborne images for citrus grove identification. ISPRS Journal of Photogrammetry and Remote Sensing 66 (1), 115–123.

Baraldi, A., Wassenaar, T., Kay, S., 2010. Operational performance of an automatic preliminary spectral rule-based decision-tree classifier of space borne very high resolution optical images. IEEE Transactions on Geoscience and Remote Sensing 48 (9), 3482–3502.

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Breiman, L., 1994. Bagging predictors, Technical Report 421. University of California, Berkeley, USA, Department of Statistics.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, CA, USA.

Brekke, C., Solberg, A., 2005. Oil spill detection by satellite remote sensing. Remote sensing of Environment 95 (1), 1–13.

Clark, M.L., Aide, T.M., Grau, H.R., Riner, G., 2010. A scalable approach to mapping annual land cover at 250 m using MODIS time series data: A case study in the Dry Chaco ecoregion of South America. Remote Sensing of Environment 114 (11), 2816–2832.

Del Frate, F., 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.

Dumas, P., Printemps, J., Mangeas, M., Luneau, G., 2010. Developing erosion models for integrated coastal zone management: a case study of The New Caledonia west coast. Marine Pollution Bulletin 61 (7–12), 519–529.

Espedal, A., Johannessen, A., 2000. Detection of oil spills near offshore installations using synthetic aperture radar (SAR). International Journal of Remote Sensing 21 (11), 2141–2144.

Fiscella, B., Giancaspro, A., Nirchio, F., Trivero, P., 2000. Oil spill detection using marine SAR images. International Journal of Remote Sensing 21 (18), 3561–3566.

Guo, L., Chehata, N., Mallet, C., Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. ISPRS Journal of Photogrammetry and Remote Sensing 66 (1), 56–66.

Ho, T., 1995. Random Decision Forest. 3rd International Conf. on Document Analysis and Recognition. 278–282.

Ho, T., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8), 832–844.

Karathanassi, V., Topouzelis, K., Pavlakis, P., Rokos, D., 2006. An object-oriented methodology to detect oil spills. International Journal of Remote Sensing 27 (23), 5235–5251.

Kavzoglu, T., Mather, P.M., 1999. Pruning artificial neural networks: an example using land cover classification of multi-sensor images. International Journal of Remote Sensing 20 (14), 2787–2803.

Keramitsoglou, I., Cartalis, C., Kiranoudis, C., 2005. Automatic identification of oil spills on satellite images. Environmental Modeling and Software 21 (5), 640–652.

Kononenko, I., Hong, J., 1997. Attribute selection for modeling. Future Generation Computer Systems 13 (2–3), 181–195.

Migliaccio, M., Trangaglia, M., 2004. Oil spill observation by SAR: a review, US-Baltic International Symposium, Klaiped, Lithuania, June 14–17.

Miller, W.F., Quattrochi, D.A., Carter, B.D., Higgs, G.K., Solomon, J.L., Wax, C.L., 1979. Application of remote sensing to state and regional problems. Semiannual progress report, 1 Nov. 1978–30 Apr. 1979.

Mingers, J., 1989. An empirical comparison of selection measures for decision tree induction. Machine Learning 3 (4), 319–342.

Montali, A., Giancito, G., Migliaccio, M., Gambardella, A., 2006. Supervised pattern classification techniques for oil spill classification in SAR images: preliminary results, SEASAR 2006 Workshop, ESA-ESRIN, Frascati, Italy, 23–26 January.

Muasher, M.J., Landgrebe, D.A., 1981. Multistage classification of multispectral earth observational data: the design approach. Purdue University, LARS Technical Reports 101481, 171.

Pavlakis, P., Tarchi, D., Sieber, A., 2001. On the Monitoring of Illicit Vessel Discharges, A reconnaissance study in the Mediterranean Sea, European Commission, EUR 19906 EN.

Pudil, P., Novovicova, J., Kittler, J., 1994. Floating search methods in feature selection. Pattern Recognition Letters 15 (11), 1119–1125.

Scholz, D., Fuhs, N., Hixson, M. and Akiyama, T., 1979. Evaluation of several schemes for classification of remotely sensed data: their parameters and performances. Purdue University, LARS Technical Report, 041279. .

Solberg, A., Brekke, C., Husoy, P.O., 2007. Oil spill detection in Radarsat and Envisat SAR images. IEEE Transactions on Geoscience and Remote Sensing 45 (3), 746–755.

Solberg, A., Storvik, G., Solberg, R., Volden, E., 1999. Automatic Detection of Oil Spills in ERS SAR Images. IEEE Transactions on Geoscience and Remote Sensing 37 (4), 1916–1924.

Solberg, R., Theophilopoulos, N.A., 1997. Envisys - A solution for Automatic oil spill detection in the Mediterranean, In: Proceedings of the Fourth Thematic Conference on Remote Sensing for Marine and Coastal Environmentsvol. 1, Environmental Research Institute of Michigan, Ann Arbor, Michigan, 3–12.

Stathakis, D., Topouzelis, K., Karathanassi, V., 2006. Large-scale feature selection using evolved neural networks, In Proceedings of SPIE, Image and Signal Processing for Remote Sensing XII, Bruzzone (Ed.), 6365.

Theodoridis, S., Koutroumbas, K., 2006. Pattern recognition, third ed. Academic Press, San Diego.

Topouzelis, K., 2008. Oil Spill Detection by SAR Images: Dark Formation Detection. Feature Extraction and Classification Algorithms, Sensors 8 (10), 6642–6659.

Topouzelis, K., Karathanassi, V., Pavlakis, P., Rokos, D., 2003. Oil Spill Detection: SAR Multi-scale Segmentation & Object Features Evaluation. In Proceedings of SPIE, Remote Sensing of the Ocean and Sea ice 2002, 23–27 September, Creta, Greece, Bostater and Santoleri (Ed.), 4880, 77–87.

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.

Yu, X., Hyyppä, J., Vastaranta, M., Holopainen, M., Viitala, R., 2011. Predicting individual tree attributes from airborne laser point clouds based on the random forests technique. ISPRS Journal of Photogrammetry and Remote Sensing 66 (1), 28–37.