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.

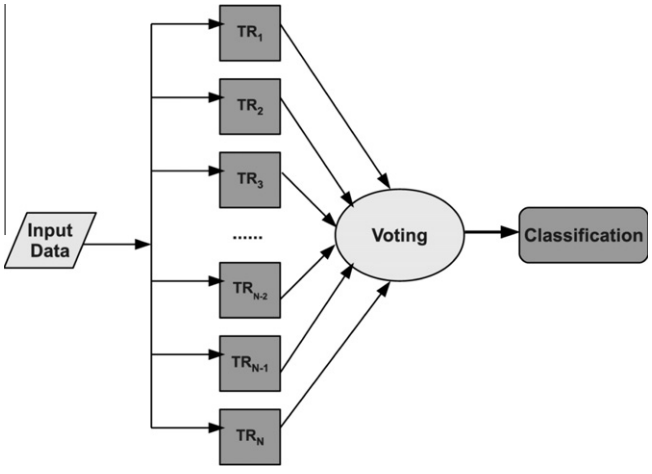

Fig. 1. Principle of decision tree classification using N decision trees (TR).

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