Ecological Informatics 42 (2017) 46–54

Contents lists available at ScienceDirect

Ecological Informatics

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

Consensus methods based on machine learning techniques for marine phytoplankton presence–absence prediction

M. Bourel[a,b], C. Crisci[c,*], A. Martínez[d]

[a] Instituto de Matemática y Estadística Prof. Ing. Rafael Laguardia, Facultad de Ingeniería, Julio Herrera y Reissig 565, CP 11200 Montevideo, Uruguay
[b] Departamento Métodos Matemático Cuantitativos, Facultad de Ciencias Económicas y Administración, Universidad de la República, Av. Gonzalo Ramírez 1926, CP 11200 Montevideo, Uruguay
[c] Centro Universitario Regional del Este, Universidad de la República, Ruta Nacional n9 y Ruta n15, CP 27000 Rocha, Uruguay
[d] Dirección Nacional de Recursos Acuáticos, M.G.A.P., Puerto de La Paloma, CP 27001 La Paloma, Rocha, Uruguay

ARTICLE INFO

Keywords:
Marine phytoplankton
Presence–absence data
Machine learning
Non-homogeneous consensus methods
Prediction

ABSTRACT

We performed different consensus methods by combining binary classifiers, mostly machine learning classifiers, with the aim to test their capability as predictive tools for the presence–absence of marine phytoplankton species. The consensus methods were constructed by considering a combination of four methods (i.e., generalized linear models, random forests, boosting and support vector machines). Six different consensus methods were analyzed by taking into account six different ways of combining single-model predictions. Some of these methods are presented here for the first time. To evaluate the performance of the models, we considered eight phytoplankton species presence–absence data sets and data related to environmental variables. Some of the analyzed species are toxic, whereas others provoke water discoloration, which can cause alarm in the population. Besides the phytoplankton data sets, we tested the models on 10 well-known open access data sets. We evaluated the models' performances over a test sample. For most (72%) of the data sets, a consensus method was the method with the lowest classification error. In particular, a consensus method that weighted single-model predictions in accordance with single-model performances (weighted average prediction error — WA-PE model) was the one that presented the lowest classification error most of the time. For the phytoplankton species, the errors of the WA-PE model were between 10% for the species *Akashiwo sanguinea* and 38% for *Dinophysis acuminata*. This study provides novel approaches to improve the prediction accuracy in species distribution studies and, in particular, in those concerning marine phytoplankton species.

1. Introduction

1.1. A brief introduction to consensus methods

In the classification framework of machine learning (ML), *ensemble methods* or *aggregating methods* consist in combining the predictions of several classifiers (also called *hypotheses* or *base classifiers*) that are performed over the same data set. The predictions are combined with the main goal of reducing variance and constructing a more stable and accurate predictor (James et al., 2014; Hastie et al., 2001; Bourel, 2012, 2013). Ensemble methods have had great success not only in the ML community, but also among researchers from different fields and with statistical modeling interests, because of their accuracy, which is generally higher than that of individual classifiers (Polikar, 2006). Despite the merits of these methods, it is often a challenge to understand completely the theoretical framework behind them.

The strategy of combining the outputs of different classifiers implies that individual classifiers make errors on different instances. The logic is that, if each classifier makes different errors, then a good combination of these classifiers can reduce the total error, improving the errors of not-so-good classifiers. For this, it is interesting to make each classifier as unique as possible with respect to misclassified instances. In particular, it is necessary to find classifiers whose decision boundaries are adequately different from those of others. Such a set of classifiers is said to be *diverse* (Polikar, 2006; Brown et al., 2005 and references therein). In general, however, ensemble algorithms do not attempt to maximize a specific diversity measure. Rather, increased diversity is usually sought somewhat heuristically through various resampling procedures, such as the selection (randomly or not) of different training parameters, models, or subsets of features.

Ensemble methods can be classified into two categories: *homogeneous* and *non-homogeneous*. Homogeneous methods combine

* Corresponding author.
E-mail addresses: mbourel@fing.edu.uy (M. Bourel), carocrisci@cure.edu.uy (C. Crisci).

http://dx.doi.org/10.1016/j.ecoinf.2017.09.004
Received 28 May 2017; Received in revised form 6 September 2017; Accepted 9 September 2017
Available online 12 September 2017

classifiers of the same nature; examples of this type of methods are *bagging* (Breiman, 1996a), *random forests (RF)* (Breiman, 2001), and *boosting* (Freund and Schapire, 1997; Schapire and Freund, 1998). In this paper, we will pay attention to non-homogeneous methods and we will refer to them as *consensus methods*. Consensus methods consist of a combination of various methods of a different nature. Examples of this type of methods are *stacking* (Wolpert, 1992; Ting and Witten, 1999; Breiman, 1996b) and *mixture of experts* (Masoudnia and Ebrahimpour, 2014). The different predictors are combined in some way; for instance, in the case of mixture of experts, this is done generally by averaging (with or without weights) or by voting over the models' predictions. In the case of stacking, the outputs of the different classifiers are used to train another classifier, which makes the final decision rule of the methods.

A way of doing a mixture of experts is inspired, to some extent, by Bayesian voting, and it consists in assigning a weight to each hypothesis (Kuncheva, 2014). A classifier h generally calculates the posterior probability that a given observation belongs to a class. To fix the notation, we can think that h computes a vector $(p_0^h(\mathbf{x}), p_1^h(\mathbf{x}))$, where $p_0^h(\mathbf{x})$ and $p_1^h(\mathbf{x})$ are the posterior probabilities that observation $\mathbf{x}$ belongs to class 0 or to class 1, respectively. The consensus of different intermediate classifiers $h_1, \ldots, h_M$ is to generate a classifier F of the form

$$F(\mathbf{x}) = \underset{k\in\{0,1\}}{\text{Argmax}} \left(\sum_{m=1}^{M} w_{h_m,\mathscr{D}}\, p_k^{h_m}(\mathbf{x}) \right).$$

This type of combination is called a weighted averaging combining rule. In this paper, we will compare it empirically to other mixture-of-expert rules and to two versions of stacking.

1.2. Consensus methods in ecological studies

Concerning the ecological modeling of species presence–absence, the performance of different statistical techniques could vary significantly from a particular case study to another, and it is not very clear sometimes which model is the most suitable. There are two possible strategies to reduce the models' uncertainty: (1) by acquiring an understanding via extensive model comparisons as to which method will generally provide the best predictive performance and in what conditions (Marmion et al., 2009b) and (2) by using consensus methods (i.e., non-homogeneous ensemble methods) (Thuiller, 2004; Thuiller et al., 2005; Araújo and New, 2007; Marmion et al., 2009b). As mentioned earlier, consensus methods overcome the problem of variability in the predictions of different single models since they are based on the combination of their predictions. Hence, a relevant combination of several unbiased (i.e., with good accuracy) model outputs will result in a more accurate prediction.

The matter rests in choosing adequate single models and finding a relevant algorithm to combine them. When dealing with ecological problems, ML techniques seem to be good candidates for single models because of their predictive capacity (Olden and Jackson, 2002). These techniques are frequently and increasingly considered in ecological studies, in particular in modeling species presence–absence or abundance from environmental variables (De'ath and Fabricius, 2000; Guisan et al., 2002; Drake et al., 2006; Cutler et al., 2007; Kampichler et al., 2010; Olden and Jackson, 2002). ML methods have advantages over traditional statistical methods (e.g., linear models and generalized linear models) since they can deal with some characteristics typical of ecological data such as unusual distributions, non-linearity, multiple missing values, complex data interactions, and dependence on the observations (Guisan et al., 2002; Cutler et al., 2007; Crisci et al., 2012). Besides their flexibility, they typically outperform traditional approaches, making them ideal for modeling ecological systems (Olden et al., 2008). In fact, concerning ecological studies, ML methods are always considered when performing consensus models (Marmion et al., 2009a,b; Lauzeral et al., 2015; Comte and Grenouillet, 2013; Thuiller et al., 2009). Besides ML techniques, more classical techniques such as generalized linear modeling or linear discriminant analysis are usually considered in the consensus construction (Thuiller et al., 2009; Marmion et al., 2009a,b; Lauzeral et al., 2015; Comte and Grenouillet, 2013) since, in some cases (e.g., linear relations between the predictors and the response variable), these methods may outperform ML techniques.

It must be noted that, although the consensus approach clearly has a number of attractive characteristics, the understanding of its merits for ecological prediction is still limited (Marmion et al., 2009b); hence, further studies comparing the predictive capacity of consensus methods with that of single methods are needed. It must be noted also that most of the applications of consensus methods in ecological studies are related to the study of species distribution models (SDMs) (Guisan and Thuiller, 2005).

In this paper, we explore the performance of six different consensus methods for predicting the presence–absence of eight marine phytoplankton species from the Atlantic coast of Uruguay. Four of the methods are a mixture of experts, and the other two are stacking applications. Moreover, we analyze the performance of the consensus models by considering 10 well-known open access data sets. To generate the consensus, we combined four individual models with very different structures, three of which have been documented as some of the most accurate ML techniques: boosting, RF, and support vector machine (SVM), whereas the fourth is a generalized linear model (GLM) that could better capture the linear relationships in data. For a more detailed description of these models, we refer the reader to the Supplementary material.

2. Methods

In this section, we present *i*) the data sets used to evaluate the performance of the models; *ii*) the principal concepts of supervised classification, *iii*) a description of the consensus models analyzed in this work; *iv*) the way in which we calculated the prediction error of the models; and *v*) the model tuning and optimization, and the use of software and functions.

2.1. Data sets

2.1.1. Marine phytoplankton data

The marine phytoplankton data set is part of the Harmful algal blooms (HABs) monitoring program, which is conducted by the National Direction of Aquatic Resources of Uruguay. The program is carried out weekly since 1991 at fixed sites in the Atlantic coast of Uruguay. We decided to consider the 2011–2014 period because data were available for a greater number of phytoplankton species; furthermore, there was more information concerning the predictor variables. For the period considered, 196 observations were available. Surveys were carried out in two exposed sandy beaches with contrasting morphodynamics: Barra del Chuy (33° 45′ S, 53° 27′ W), which is a dissipative beach with fine to very fine well-stored sand, a gentle slope, heavy wave action, and a wide surf zone; and Arachania (34° 36′, 53° 44′ W), which is a reflective beach with coarse sediments and a steep slope (Bergamino et al., 2016) (Fig. 1). At each site, water samples were taken from the surf zone with a plastic bucket for chlorophyll *a* and phytoplankton quantification. Moreover, water temperature and salinity were measured *in situ* with an ISY *ECO*300 probe, and wind intensity and direction were estimated visually. Phytoplankton species were identified and counted in an Olympus IM inverted microscope thereafter Utermöhl (1958) at a final magnification of 1000 × (Andersen and Throndsen, 2003). Furthermore, the abundance of potential phytoplankton consumers was registered. Because of potential differences in prey preferences, we decided to consider the three following guilds of phytoplankton consumers: *i*) microcrustaceans, *ii*) ciliates and tintinids, and *iii*) ciliates, tintinids, and heterotrophic