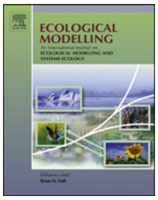
Contents lists available at SciVerse ScienceDirect

Ecological Modelling

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



Review

A review of supervised machine learning algorithms and their applications to ecological data

C. Crisci[a], B. Ghattas[b,*], G. Perera[c]

[a] UMR, 6540, CNRS, Université de la Méditerranée, DIMAR, Centre d'Océanologie de Marseille, Station Marine d'Endoume, Chemin de la Batterie des Lions, Marseille 13007, France
[b] Université de la Méditerranée, Département de Mathématiques, Case 901, 163 avenue de Luminy, Marseille 13009, France
[c] Facultad de Ingenieria de la Universidad de la República, Montevideo, Uruguay

ARTICLE INFO

Article history:
Received 11 July 2011
Received in revised form 8 January 2012
Accepted 1 March 2012
Available online 16 June 2012

Keywords:
Machine learning
Ecological data
Regression analysis
Classification rules
Prediction
Mass mortality events
Coastal rocky benthic communities
Positive thermal anomalies

ABSTRACT

In this paper we present a general overview of several supervised machine learning (ML) algorithms and illustrate their use for the prediction of mass mortality events in the coastal rocky benthic communities of the NW Mediterranean Sea. In the first part of the paper we present, in a conceptual way, the general framework of ML and explain the basis of the underlying theory. In the second part we describe some outstanding ML techniques to treat ecological data. In the third part we present our ecological problem and we illustrate exposed ML techniques with our data. Finally, we briefly summarize some extensions of several methods for multi-class output prediction.

Contents

* Corresponding author.
E-mail addresses: carolina.crisci@univmed.fr (C. Crisci), ghattas@univmed.fr (B. Ghattas), gperera@fing.edu.uy (G. Perera).

0304-3800/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2012.03.001

1. Introduction

Ecological systems rarely require simple statistical analysis. For example, much of the data collected by ecologists often exhibit *unusual* distributions, non-linearity, multiple missing values, complex data interactions, dependence on the observations, etc. (Fielding, 1999; De'ath, 2007; Guisan et al., 2002; Cutler et al., 2007). The size of datasets is another very hard and frequent problem. Many of the ecological databases are very large and some are continuously expanding. The problems related to a large number of cases or variables are likely to become more severe as more biodiversity data are accumulated and remotely sensored data are increasingly used (Fielding, 1999). Bellman's "curse of dimensionality", often invoked by statisticians to refer to how difficult it may be to deal with a huge amount of data, appears nowadays frequently in Ecology.

Machine learning (ML) techniques are not and will never be the solution to all the problems risen by ecological data. However, these techniques provide a powerful set of tools that deserves a serious attention to deal with some relevant ecological problems. As a first point, let us remark that ML concerns are nothing but the most ancient, classical and widely studied statistical problems: classification, regression, decision, clustering, density estimation, etc. However, what makes ML a particular field is not precisely its goals and problems but its tools, techniques and strategies, characterized by the massive use of algorithms and computational resources to deal with large sets of data, high number of variables and complex data structures.

ML approaches are intensively applied in different areas and there is no doubt that Ecology is today one of the most relevant areas of ML application (Flach, 2001). This is reflected in the large number of publications that appeared in the last years in which diverse ML techniques are applied to solve a wide variety of problems. Studies of the relationship between organisms (species presence/absence, population and community attributes) and habitat characteristics applying ML techniques are well documented in terrestrial (e.g. Ryder and Irwin, 1987; Franklin, 1998; Shan et al., 2006; Cutler et al., 2007), fresh water (e.g. Lek and Guégan, 1999; Džeroski, 2001; Kocev et al., 2010) and marine ecosystems (e.g. De'ath and Fabricius, 2000; Defeo and Gómez, 2005; Merckx et al., 2009; Knudby et al., 2010; Volf et al., 2011). Some particular applications of these techniques in Ecology are the prediction of algal blooms (Ribeiro and Torgo, 2008), fish recruitment (Fernandes et al., 2010), habitat suitability of tree species (Benito Garzón et al., 2006), organism identification (Morris et al., 2001) and determination of factors affecting dispersal of marine species (Pontin et al., 2011).

Previous reviews of ML methods expose in detail a few number of techniques (Recknagel, 2001). Some of them illustrate some ML techniques through case studies but without expanding on theoretical basis (Džeroski, 2001). Further articles present the theoretical basis of some specific technique in a more or less formal manner, illustrating them with ecological examples (Lek and Guégan, 1999; De'ath and Fabricius, 2000; Guisan et al., 2002; De'ath, 2007).

Here we intend to present a comprehensive view of ML techniques giving a brief overview of the theory underlying these approaches. We restrict our study to situations where we wish to model the effect of a set of explanatory variables (X) on a target variable (Y). This is the context of *Supervised Learning* (SL) said otherwise, *Regression* or *Classification* depending on the nature of Y. We have selected eight methods among the large panel of available approaches. Theses methods gave rise since the 1995th to extensive research in the machine learning community as they have numerous advantages among which being non-parametric and free from any distribution assumption. Besides, most of these methods offer a lot of extensions and may be used to model multidimensional outputs or functional outputs. Finally, they may be mathematically unified, as they be expressed as linear or convex combinations of non-parametric functions.

We illustrate these methods with an ecological example, the prediction of mass mortality events in the NW Mediterranean coastal rocky benthic communities.

The paper is organized as follows. In Section 2 the general framework of SL is presented with a glance of its theoretical basis but explained in a conceptual manner. In Section 3 we present a wide panel of SL techniques and in Section 4 we illustrate the exposed techniques with our application. Finally, in Section 5 we summarize some technical extensions useful in particular ecological problems such as prediction of multi-class and functional outputs.

2. Supervised Learning (SL): general framework and fundamentals

The main purpose of SL techniques is to learn how to predict a random variable $Y \in \mathcal{Y}$ based on a set of explicative random variables denoted by $X \in \mathcal{X}$, where $\mathcal{Y}$ and $\mathcal{X}$ depend on the problem at hand but may be thought to be respectively $\mathbb{R}$ and $\mathbb{R}^d$ for example. We will often call X the *input* and Y the *output*. As a leading example, one may think about a variable Y that represents the presence/absence of a rare lichen species (Cutler et al., 2007), and a set of variables X that consists on elevation, aspect and slope. The main problem is to find a *predictor*:

$$f: \quad \mathcal{X} \to \mathcal{Y}$$

$$X \to f(X)$$

chosen among the set of all functions $\mathcal{F} = \{f: \ \mathcal{X} \to \mathcal{Y}\}$. To build a "good" predictor we have to define a performance criterion that is, a *loss function* denoted L which depends namely on f, X and Y. We thus say that predictor f is better than predictor g if $L(f, X, Y) < L(g, X, Y)$. To simplify the notation we will omit the dependence of L on X and Y.

Suppose that there exists a unique predictor $f^* \in \mathcal{F}$ which minimizes the loss function L, called the *optimal predictor*,

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, L(f, X, Y).$$

In general however, it is not possible to minimize L over the whole set of possible functions $\mathcal{F}$ (that may be a very large set) but only over a given class of predictors C that corresponds to a set of practically computable predictors (for instance the class of linear models). In such case one obtains a predictor f^{**} satisfying

$$f^{**} = \underset{f \in \mathcal{C}}{\operatorname{argmin}}\, L(f, X, Y).$$

The predictor f^{**} may be different of the globally optimal predictor f^* (which may be not included in C). Besides, the predictor