Ecological Modelling 221 (2010) 330–337

Contents lists available at ScienceDirect

Ecological Modelling

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

Learning habitat models for the diatom community in Lake Prespa

Dragi Kocev[a], Andreja Naumoski[b], Kosta Mitreski[b], Svetislav Krstić[c], Sašo Džeroski[a,*]

[a] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[b] Department of Computer Technologies and Environment Centre, Faculty of Electrical Engineering and Information Technology, Skopje, Macedonia
[c] Institute of Biology, Faculty of Natural Sciences and Mathematics, Skopje, Macedonia

ARTICLE INFO

Article history:
Received 15 May 2008
Received in revised form 26 August 2009
Accepted 4 September 2009
Available online 12 October 2009

Keywords:
Diatom community
Habitat modelling
Multi-target modelling
Regression trees
Lake Prespa

ABSTRACT

Habitat suitability modelling studies the influence of abiotic factors on the abundance or diversity of a given taxonomic group of organisms. In this work, we investigate the effect of the environmental conditions of Lake Prespa (Republic of Macedonia) on diatom communities. The data contain measurements of physical and chemical properties of the environment as well as the relative abundances of 116 diatom taxa. In addition, we create a separate dataset that contains information only about the top 10 most abundant diatoms. We use two machine learning techniques to model the data: regression trees and multi-target regression trees. We learn a regression tree for each taxon separately (from the top 10 most abundant) to identify the environmental conditions that influence the abundance of the given diatom taxon. We learn two multi-target regression trees: one for modelling the complete community and the other for the top 10 most abundant diatoms. The multi-target regression trees approach is able to detect the conditions that affect the structure of a diatom community (as compared to other approaches that can model only a single target variable). We interpret and compare the obtained models. The models present knowledge about the influence of metallic ions and nutrients on the structure of the diatom community, which is consistent with, but further extends existing expert knowledge.

1. Introduction

Ecology is frequently defined as the study of the distributions and abundances of organisms across space and time and their interactions with the environment (Begon et al., 2006). Habitat modelling focuses on the spatial aspects of the distribution and abundance of plants and animals. It studies the relationships between some environmental variables and the presence/abundance of plants and animals. This is typically done under the implicit assumption that both are observed at a single point in time for a given spatial unit.

The input to a habitat model (Džeroski, 2001, 2009) is a set of environmental characteristics for a given spatial unit of analysis. These environmental characteristics (i.e., environmental variables) may be of three different types. The first type concerns abiotic properties of the environment, e.g., physical and chemical characteristic thereof. The second type concerns some biological aspects of the environment, which may be considered as an external impact on the group of organisms under study. Finally, the variables of the third type are related to human activities and their impacts on the environment. The output of a habitat model is a target property of the given (taxonomic) group of organisms. Note that the type of environmental variables, as well as the size of the spatial unit, can vary considerably, depending on the context, and so can the target property of the population (even though to a lesser extent). If we take the abundance or density of the population as indicators of the suitability of the environment for the group of organisms studied, we talk about habitat suitability models: the output of these models can be interpreted as a degree of suitability. The abundance of the population can be measured in terms of the number of individuals or their total size (e.g., the dry biomass of a certain species of algae). If the (taxonomic) group is large enough, we can also consider the diversity of the group (e.g., Shannon index, species richness).

In the most general case of habitat modelling, we are interested in the relation between the environmental variables and the structure of the population at the spatial unit of analysis (absolute and relative abundances of the organisms in the group studied). One approach to this is to build habitat models for each of the organisms (or lower taxonomic units) in the group, then aggregate the outputs of these models to determine the structure of the population. An alternative approach is to build a model that simultaneously predicts the presence/abundance of all organisms in the group.

In this work, we explore the two afore mentioned possibilities for habitat modelling of the diatom community in Lake Prespa

* Corresponding author.
E-mail addresses: Dragi.Kocev@ijs.si (D. Kocev), Andreja.Naumoski@feit.ukim.edu.mk (A. Naumoski), komit@feit.ukim.edu.mk (K. Mitreski), skrstic@iunona.pmf.ukim.edu.mk (S. Krstić), Saso.Dzeroski@ijs.si (S. Džeroski).

0304-3800/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.09.002

(Republic of Macedonia). To learn a model for each diatom taxon separately, we employ regression trees (Breiman et al., 1984). To build a model for the entire diatom community, we use multi-target regression trees (Blockeel et al., 1998; Struyf and Džeroski, 2006). The main advantages of the latter approach are: (1) the multi-target model is smaller and faster to learn than learning models for each organism separately and (2) the dependencies between the organisms are explicated and explained.

The data that we use were collected during the EU funded project TRABOREMA (FP6-INCO-CT-2004-509177). They describe the diatom abundance in Lake Prespa. The measurements comprise several important parameters that reflect the physical, chemical and biological aspects of the water quality of the lake. These include measurements of the relative abundance of algal taxa belonging to the group *Bacillariophyta* (diatoms). The focus of this paper is the investigation of the relationship between their relative abundance and the abiotic characteristics of the environment (Lake Prespa).

Diatoms have narrow tolerance ranges for many environmental variables and respond rapidly to environmental change. This makes them ideal bio-indicators (Reid et al., 1995; Round, 1991). They are sensitive to changes in nutrient concentrations, supply rates and silica/phosphate ratios; they respond rapidly to eutrophication. Each taxon has a specific optimum and tolerance for nutrients such as phosphorus and nitrogen. Diatoms are widely used as bio-indicators in Europe (Krstić, 1995; Krstić et al., 1998; Krstić et al., 2007; Kelly et al., 1998; Prygiel and Coste, 1999), North America (Stevenson and Pan, 1999; Lowe and Pan, 1996), South America (Lobo et al., 1998; Loez and Topalian, 1999) and Australia (John, 1998; Chessman et al., 1999). The geographical location of the diatoms is not the limiting factor in the distribution of diatom taxa and the composition of communities; rather, the specific environmental variables prevailing at a particular location (Gold et al., 2002) are the limiting factors.

The remainder of this paper is organized as follows. In Section 2, we describe the machine learning methodology that was used (regression trees and multi-target regression trees). Section 3 describes the data and Section 4 explains the experimental design that was employed to analyze the data at hand. Section 5 presents the obtained models and discusses them, while Section 6 concludes.

2. Machine learning for habitat modelling

2.1. Machine learning basics

The input to a machine learning algorithm is most commonly a single table of data comprising a number of fields (columns) and records (rows) (Džeroski, 2001, 2009). In general, each row represents an object and each column represents a property (of the object). In machine learning terminology, rows are called examples and columns are called attributes (or sometimes features). Attributes that have numeric (real) values are called continuous attributes. Attributes that have nominal values are called discrete attributes.

The tasks of classification and regression are the two most commonly addressed tasks in machine learning. They are concerned with predicting the value of one field from the values of other fields. The target field is called the target attribute or class (dependent variable in statistical terminology). The other fields are called descriptive attributes or just attributes (independent variables in statistical terminology). If the class is continuous, the task at hand is called regression. If the class is discrete (it has a finite set of nominal values), the task at hand is called classification. In both cases, a set of data (dataset) is taken as input, and a predictive model is generated. This model can then be used to predict values of the class for new data.

To estimate the performance of the model on unseen data, several approaches can be used (Kohavi, 1995). One approach consists of dividing the data in two parts (typically 2/3 and 1/3): training set (the bigger part) and testing set (smaller part). The most commonly used approach is cross-validation. The division into a training/testing set is recommended in the case of datasets that contain many records (thousands); cross-validation is a better choice otherwise.

2.2. A machine learning formulation of the habitat modelling task

In the case of habitat modelling, examples correspond to spatial units of analysis. The attributes correspond to environmental variables describing the spatial units, as these are the inputs to a habitat model. The class is a target property of the given (taxonomic) group of organisms, such as presence, abundance or diversity.

The machine learning task of habitat modelling (Džeroski, 2009) is thus defined as follows. Given is a set of data with rows corresponding to spatial locations (units of analysis), attributes corresponding to environmental variables, and the class corresponding to a target property of the population studied. The goal is to learn a predictive model that predicts the target property from the environmental variables (from the given dataset). If we are only looking at presence/absence or suitable/unsuitable as values of the class (as is the case above), we have a classification problem. If we are looking at the degree of suitability (density/abundance), we have a regression problem.

2.3. Regression trees

Regression trees are decision trees that are capable of predicting the value of a numeric target variable (Breiman et al., 1984). They are hierarchical structures, where the internal nodes contain tests on the input attributes. Each branch of an internal test corresponds to an outcome of the test, and the predictions for the values of the target attribute are stored in the leaves. Regression tree leaves contain constant values as predictions for the target variable (they represent piece-wise constant functions).

To obtain the prediction of a regression tree for a new data record, the record is sorted down the tree, starting from the root (the top-most node of the tree). For each internal node that is encountered on the path, the test that is stored in the node is applied, and depending on the outcome of the test, the path continues along the corresponding branch (to the corresponding subtree). The procedure is repeated until we end up in a leaf. The resulting prediction of the tree is taken from this leaf.

The tests in the internal nodes can have more than two outcomes (this is usually the case when the test is on a discrete-valued attribute, where a separate branch/subtree is created for each value). Typically, each test has two outcomes: the test has succeeded or the test has failed. The trees in this case are called binary trees.

2.4. Multi-target regression trees

Multi-target regression trees are an instantiation of predictive clustering trees (PCTs) (Blockeel et al., 1998), where a tree is viewed as a hierarchy of clusters. The top-node of a PCT corresponds to a cluster that contains all the data. This cluster is then recursively partitioned into smaller clusters while moving down the tree. The leaves represent the clusters at the lowest level of the hierarchy and each leaf is labelled with its prototype.

Multi-target regression trees (Blockeel et al., 1998; Struyf and Džeroski, 2006) are a generalization of regression trees, because they can predict the values of several numeric target attributes simultaneously. Instead of storing a single numeric value, the leaves