A Comparative Study on Predicting Algae Blooms in Douro River, Portugal

Rita Ribeiro[1], Luís Torgo[2]
[1]ILIACC-University of Porto, R. Ceuta 118 – 6o., 4050-190 Porto, Portugal, rita@liacc.up.pt,
Phone: +351 22 339 2090, Fax: +351 22 339 2099
[2]ILIACC/FEP, University of Porto, R. Ceuta 118 – 6o., 4050-190 Porto, Portugal,
ltorgo@liacc.up.pt, Phone: +351 22 339 2093

Key words: Multiple regression, extreme values prediction, model evaluation, algae blooms.

Introduction

This paper describes the results of a comparative experiment using different models for predicting the abundance of different groups of phytoplankton in water samples collected in Douro River, Porto, Portugal. Several data are collected at the Crestuma-Lever dam in river Douro, with the goal of controlling water quality, as this is the source of potable water of the metropolitan area of Porto, the second largest city in Portugal. These include identification and quantification of different groups of phytoplankton, physico-chemical parameters analyses as well as microbiological analysis. The original data used in this study was obtained from the public company responsible for potable water collection. The data spans the period from 1998 till 2003. Unfortunately, the periodicity and the factors analysed throught this period are far from being homogenius. This required large pre-processing steps to came up with a dataset useful for model construction. The long term objective of the project behind this paper is to develop reliable models that can forecast microalgae blooms in river Douro. Identification of of the different groups of phytoplankton in water samples requires intensive and expensive manual labour, while the analysis of most physico-chemical parameters can be automated by water probes. Our objective is to avoid the expensive manual identification of phytoplankton groups by using models that are able to accurately predict the abundance of certain groups of microalgae based on the values of the certain physico-chemical parameters. In this paper we present the results of a first step towards this goal. Namely, we describe the results of a comparative experiment between several candidate models. Predicting the abundance level of a certain phytoplankton species given the values of a set of physico-chemical parameters can be regarded as a multiple regression problem. Several techniques exist for handling this type of problems. They all revolve around the issue of obtaining the model parameters based on a data sample, which optimize a certain preference criteria. A typical preference criterion is the mean squared error (MSE) of the model predictions. While this type of criteria are adequate for most applications, we claim that they introduce a wrong bias for this particular application. In effect, accurate anticipation of microalgae blooms is the key objective of our task. Microalgae blooms are fortunately rare events. As such, the distribution function of the abundance of a certain groups of phytoplankton has a normal shape but with unusually large values (the blooms) occurring with very low frequency. Given the stated objectives of our application we should prefer a model that is accurate at these rare cases and less accurate at the most frequent (non-bloom) cases, to a model that drastically fails the prediction of the blooms whilst being on average very accurate at the most frequent values of the target variable (the abundance). However, error statistics like the MSE will tend to favour the model that has good performance on the most frequent cases. As such, the use of such preference criteria will introduce a bias into the obtained models that goes against the goals of the application, i.e. predicting microalgae blooms. Based on these observations, Torgo and Ribeiro (2005) have developed a different error statistic that is able to identify the models that are more "useful" from the perspective of accurately predicting the rare extreme values of the target variable. In our comparative experiments we have used this statistic to compare the models we have tried in our data.

Results and discussion

We have carried out a Monte Carlo experimental comparison between different types of models on the task of predicting the abundance of several harmful microalgae. The models included several variants of regression trees (Breiman et al., 1984), support vector machines (Vapnik, 1995) and neural networks (Rumelhart et al., 1986). All model variants were obtained using two years of data and tested on the subsequent 9 months. We have used the error statistic describe in Torgo and Ribeiro (2005) to evaluate the ability of the different models to accurately predict the rare extreme values of the algae abundances. The results of our experiments have shown that one of the used variants of a neural network clearly outperforms the other models that were tried. The results of our experiments provide indications of the wrong bias of most used models in terms of predicting rare extreme values.

Conclusions

This paper presents a comparative study of different prediction models on the difficult task of predicting algae blooms using physico-chemical parameters of water samples in river Douro, Portugal. Our study confirms the difficulty of the problem caused mainly by the scarcity of the phenomena we want to accurately predict. This type of rare events require specific evaluation criteria. We have used on of such criteria to evaluate the performance of several models. Our results indicate a variant of a neural network as the most promising model. Future work will include the modification of the modelling techniques so as to make them obtain the model parameters that are better according to this new preference criteria.

References

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and regression trees.* Wadsworth, 1984.

Rumelhart,D., Hinton,G., and Williams,R. Learning internal representations by back propagation. MIT Press, 1986.

Torgo,L. and Ribeiro,R. : Predicting Rare Extreme Values. Submitted. 2005.

Vapnik, V. *The Nature of Statistical Learning Theory*. Springer, 1995.