Available at www.sciencedirect.com

ScienceDirect

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

Robust predictive modelling of water pollution using biomarker data

Marcin Budka [a,*], *Bogdan Gabrys* [a], *Elisa Ravagnan* [b]

[a] *Computational Intelligence Research Group, Smart Technology Research Centre, School of DEC, Bournemouth University, Poole House, Talbot Campus, Fern Barrow, Poole BH12 5BB, United Kingdom*

[b] *International Research Institute of Stavanger, Mekjarvik 12, 4070 Randaberg, Norway*

ARTICLE INFO

Article history:
Received 23 October 2009
Received in revised form
22 February 2010
Accepted 3 March 2010
Available online 16 March 2010

Keywords:
Biomarkers
Water quality monitoring
Marine pollution
Ensemble classification
Missing data
Predictive modelling

ABSTRACT

This paper describes the methodology of building a predictive model for the purpose of marine pollution monitoring, based on low quality biomarker data. A step-by-step, systematic data analysis approach is presented, resulting in design of a purely data-driven model, able to accurately discriminate between various coastal water pollution levels.

The environmental scientists often try to apply various machine learning techniques to their data without much success, mostly because of the lack of experience with different methods and required 'under the hood' knowledge. Thus this paper is a result of a collaboration between the machine learning and environmental science communities, presenting a predictive model development workflow, as well as discussing and addressing potential pitfalls and difficulties.

The novelty of the modelling approach presented lays in successful application of machine learning techniques to high dimensional, incomplete biomarker data, which to our knowledge has not been done before and is the result of close collaboration between machine learning and environmental science communities.

1. Introduction

Water pollution monitoring becomes a crucial problem as more and more contaminants enter the marine environment every year Livingstone et al. (2000). The current trend is prediction of the toxicity level using various measurable attributes of the aquatic environment Pace (2001). This can be observed by a worldwide increase in the number of water quality research funding opportunities by the European Commission[1], the National Research Council in Canada and USA[2,3] and various local Councils. The data used in this research has been collected as a part of the 'Marine Environment IQ' project[4] funded by the Research Council of Norway[5], which run between 2006 and 2008.

The condition of a marine environment not always can be diagnosed by chemical analysis of the water, as it does not

* *Corresponding author*. Tel.: +44 1202 524111x61463; fax: +44 1202 962736.
E-mail addresses: mbudka@bournemouth.ac.uk (M. Budka), bgabrys@bournemouth.ac.uk (B. Gabrys), elisa.ravagnan@iris.no (E. Ravagnan).

[1] European Commission Research, http://ec.europa.eu/research/index.cfm
[2] National Research Council Canada, http://www.nrc-cnrc.gc.ca/eng/index.html
[3] National Research Council, http://sites.nationalacademies.org/NRC/index.htm
[4] Developing an Index of the Quality of the Marine Environment (Marine Environment IQ) based on biomarkers: Integration of pollutant effects on marine organisms, http://www.iris.no/Internet/NFR-feb2009.nsf/
[5] Research Council of Norway, http://www.forskningsradet.no/

0043-1354/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.watres.2010.03.006

provide any information regarding the health of the organisms. Moreover it may also fail to detect any pollution at all due to its low, yet biologically significant degree or very slow increase of contamination level. The solution to this problem is the use of biomarkers.

For many years biomarkers have been successfully used as a tool of exposure analysis. Their importance results from the fact, that they enable detection of pollutants not possible to achieve by other, commonly used methods like chemical or physical analysis (Ott et al., 2006; Peakall, 1994). A large number of biomarkers related to their potential effect on organisms has been developed in the literature (Depledge and Fossi, 1994; Depledge et al., 1995; Harvey and Parry, 1997; Regoli et al., 1998; Bresler et al., 2003; Hellou and Law, 2003; Rank and Jensen, 2003; Barsiene et al., 2004; Dahlhoff, 2004; Moore et al., 2004; Yang et al., 2004; Amiard et al., 2006; Bocchetti and Regoli, 2006; Lesser, 2006; Magni et al., 2006; Widdows and Staff, 2006). Although biomarkers play a great role in ecotoxicology and environmental risk assessment, they are sometimes difficult to interpret. It is problematic to determine whether a biomarker response is an indicator of impairment or is a part of the homeostatic response, indicating that an organism is successfully dealing with the exposure (Forbes et al., 2006). When dealing with mixtures of pollutants, a group of biomarkers ('battery') is usually used (Eason and O'Halloran, 2002; Chèvre et al., 2003), combining the effect and exposure tests. One of the objectives of this study was to validate the choice of biomarkers made during the 'Marine Environment IQ' project.

The acquisition of biomarker data is an involved process, which requires performing a set of usually destructive tests on biological material. The indicator species of choice are often mussels, which have been used as sentinel organisms from the 1970s (Goldberg, 1986; Goldberg and Bertine, 2000). There are multiple advantages of using bivalves in environmental monitoring as they are widely distributed, sedentary and easy to sample, they tolerate a wide range of environmental conditions and bioconcentrate environmental toxicants due to their high filtration activity. Unfortunately, in the majority of the studies it is impossible to use the same animal for the whole battery of test, because of the quantity of biological material required to perform chemical analyses (especially when using small animals like mussels). This dramatically reduces the quality of data by introducing missing attribute values and can have even more serious consequences. It is a common practice to pair the samples in order to have enough material to perform the chemical tests. This can however change the statistical properties of the data, leading to unexpected behavior of developed models, including false, highly positively biased accuracy estimates, which in consequence renders the models useless.

After the data has been collected it can finally be processed, which is the main focus of this paper. Although there have been several approaches to water quality prediction in the literature using neural networks (Maier and Dandy), self organizing maps (Aguilera et al., 2001), Bayes networks (Reckhow, 1999) and other methods (Hamilton and Schladow, 1997), to our knowledge none of them was using biomarker data. From the point of view of data modelling, the biomarker data usually has low quality due to the missing values, high dimensionality and small size of the dataset, which can cause various problems (Bishop, 1995; Duda et al., 2000). Perhaps the most important of them is to define what does one expect the data to reveal and is the data adequate for this purpose. Apart from that issue, this paper addresses the choice of appropriate modelling technique from a large number of available methods, reliable estimation of future performance of the model and various ways of dealing with imperfections of data.

On many occasions researchers from outside the machine learning community try to apply various machine learning techniques to their data without much success. This frequently is a result of treating the machine learning methods as 'black boxes', while unfortunately, in most cases, successful and efficient use of these tools requires appropriate technical knowledge and experience. Thus this paper is a result of collaboration between the machine learning and environmental science communities, which shows and discusses how to design a purely data-driven, usable solution, making use of limited and deficient input data.

The rest of this paper is organized as follows. Section 2 describes the basic properties of the dataset, including some of its statistical characteristics. In Section 3 a model development workflow consisting of a number of clearly defined steps is proposed, allowing for systematic data analysis and predictive model building. In Section 4, first individual models are built and their performance is measured for a number of data usage scenarios. Section 5 deals with the feature selection problem, investigating which biomarkers to use and which of them are not relevant for the problem at hand. In Section 6, an ensemble model is described, addressing in detail each of the difficulties caused by imperfections of the data. The experimental results for the ensemble model are given in Section 7, in which it is also demonstrated how the performance improves by building a multi-stage combination of models. Finally, conclusions are given in Section 8.

2. Dataset properties

2.1. Overview

The dataset contains a collection of biomarker data measured on mussels at 4 different marine stations located in South-West Norway (Rogaland County), in the course of a 4-week experiment. The stations have been chosen according to known water pollution levels (Grøsvik et al., 1999; Eriksen and Tvedten, 2002; Tvedten et al., 2002; Tvedten, 2003; Zorita et al., 2006) and the goal of the study was to provide field data to investigate the possible biomarker combinations to discriminate between various pollution levels. There are 50 objects[6] in the dataset, each having 12 attributes[7]. There are also 5 different classes, denoting the 5 stations, and 4% of attributes are missing. The locations of the sites can be seen in Fig. 1,

[6] The words 'object', 'instance' or 'sample' are used interchangeably.

[7] The words 'attribute', 'feature' or 'biomarker' are used interchangeably.