Learning to infer chemical parameters of river water quality from bioindicator data

Sašo Džeroski, Damjan Demšar
Department of Intelligent Systems
Jožef Stefan Institute
Jamova 39, SI-1111 Ljubljana, Slovenia
{saso.dzeroski,damjan.demsar}@ijs.is

Jasna Grbović
Hydrometeorological Institute
Vojkova 1b, SI-1000 Ljubljana, Slovenia
jasna.grbovic@rzs-hm.si

William J. Walley
School of Computing
Staffordshire University
Beaconside, Stafford ST18 0DG, UK
w.j.walley@soc.staffs.ac.uk

Abstract

We address the problem of inferring chemical parameters of river water quality from biological ones. This task is important for enabling selective chemical monitoring of river water quality. We apply machine learning, in particular regression tree induction, to biological and chemical data on the water quality of Slovenian rivers. Regression trees are constructed that predict values of chemical parameters from data on the presence of bioindicators.

1 Introduction

The quality of surface waters, including rivers, depends on their physical, chemical and biological properties. The latter are reflected by the types of living organisms present in the water and their density (this includes the structure of the community and its diversity). Based on the above properties, surface waters are classified into (one of) several quality classes which indicate the suitability of the water for different kinds of use.

Since Kolkwitz and Marsson (1902) first proposed the use of biota as a means of monitoring the quality of natural waters, many different methods for mapping biological data to discrete quality classes or continuous scales have been developed (for an overview, see De Pauw and Hawkes 1993). Most of these approaches use indicator organisms (bioindicators), which have well known ecological requirements and are selected for their sensitivity / tolerance to various kinds of pollution. Given a biological sample, information on the presence and density of all indicator organisms present in the sample is usually combined to derive a biological index that reflects the quality of the water at the site where the sample was taken.

Bioindicators can be identified at different taxonomical levels, e.g., at the species level or the family level. A family, a species or any other taxonomical group can be referred to as a taxon (plural taxa). In the Saprobic System (Kolkwitz and Marsson 1902), bioindicators are identified at the species level, which is more demanding in terms of sample processing effort, but also gives a more precise picture of the water quality. Family level identification is used in the Biological Monitoring Working Party Score (ISO-BMWP 1979), abbreviated as BMWP, and its derivative Average Score Per Taxon (ASPT), which are used in the United Kingdom.

It is well known that the physical and chemical properties give a limited picture of water quality at a particular point in time, while the biota (living organisms) act as continuous monitors of water quality over a period of time (Cairns et al. 1968). This has increased the importance of biological methods for monitoring water quality (De Pauw and Hawkes 1993) to such an extent that it has almost replaced chemical monitoring in some countries, such as the United Kingdom.

The relation between biological and chemical parameters of river water quality is an important and largely open research topic. We have already applied machine learning to the task of inferring biological parameters from chemical ones by learning rules that predict the presence of individual bioindicator taxa from the values of chemical measurements (Džeroski and Grbović 1995). In this paper, we address the

ERK'97, Portorož, B: 129-132

task of inferring chemical parameters from biological ones by predicting the values of individual chemical parameters from data on the presence of bioindicator taxa. To this end, we apply the machine learning approach of regression tree induction.

The problem of inferring chemical parameters from biological ones is practically relevant, especially in countries where extensive biological monitoring is conducted, but few chemical parameters are measured regularly. This is, for example, the case in the United Kingdom, where chemical parameters that are monitored regularly comprise basically Biological Oxygen Demand (BOD), Dissolved Oxygen (DO), Ammonia (NH_4) and Alkalinity (pH). Biological samples may, for example, reflect an increase in pollution and indicate likely causes (chemical parameters) or sources of pollution.

The remainder of the paper is organized as follows. Section 2 describes the biological and chemical data on the water quality of Slovenian rivers, as well as the setup of the experiments. The results of the experiments are presented in Section 3. Section 4 concludes with a brief discussion.

2 Data and experiments

The data

The data about Slovenian rivers come from the Hydrometeorological Institute of Slovenia (Hidrometereološki Zavod Republike Slovenije, abbreviated as HMZ) that performs water quality monitoring for most Slovenian rivers and maintains a database of water quality samples. The data provided by HMZ cover a six year period, from 1990 to 1995. Biological samples are taken twice a year, once in summer and once in winter, while physical and chemical analyses are performed several times a year for each sampling site. The physical and chemical samples include the measured values of 16 different parameters: biological oxygen demand (BOD), chlorine concentration (Cl), CO_2 concentration, electrical conductivity, chemical oxygen demand ($K_2Cr_2O_7$ and $KMnO_4$), concentrations of ammonia (NH_4), NO_2, NO_3 and dissolved oxygen (O_2), alkalinity (pH), PO_4, oxygen saturation, SiO_2, water temperature, and total hardness.

The biological samples include a list of all taxa present at the sampling site and their density. The frequency of occurrence (density) of each present taxon is recorded by an expert biologist at three different qualitative levels, where 1 means the taxon occurs incidentally, 3-frequently, and 5-abundantly. In total, 1061 water samples were available on which both physical/chemical and biological analyses were performed: our experiments were conducted using these samples.

Table 1: Correlations between actual parameter values and values predicted with regression trees. LM refers to linear models in the leaves.

Parameter predicted	r (with LM)	r (without LM)
BOD	0.659	0.652
Cl	0.590	0.578
CO_2	0.426	0.398
Electrical conductivity	0.581	0.568
$K_2Cr_2O_7$	0.624	0.602
$KMnO_4$	0.558	0.542
NH_4	0.647	0.664
NO_2	0.387	0.373
NO_3	0.406	0.378
O_2	0.550	0.544
Alkalinity (pH)	0.409	0.421
PO_4	0.457	0.461
Oxygen saturation	0.486	0.462
SiO_2	0.484	0.490
Water temperature	0.595	0.574
Total hardness	0.560	0.539

The experiments

Approximately 850 different taxa appear in the biological samples. We restricted our attention to the 415 taxa that appear in at least 10 samples. These were used as attributes (independent variables), while each of the 16 physical and chemical parameters was used as a class (dependent variable). In this way, 16 different learning problems were formulated.

As the physical and chemical parameters are real-valued, we used the system M5.1 (Quinlan 1993) to induce regression trees for each of the 16 problems. Ordinary regression trees and trees with linear models in the leaves were considered. Otherwise, the default parameters of M5.1 were used in all experiments. For each problem, the following methodology was employed. First, construct a single regression tree from the entire dataset. To estimate the performance of the tree on unseen data, conduct 10-fold cross-validation. These results are given in Table 1. Finally, show the tree to domain experts to verify that it contains useful domain knowledge.

3 Results

The correlation between the values of parameters predicted by the induced trees and the real parameter values is given in Table 1: the second column lists correlation for trees with linear models in leaves, the third for ordinary regression trees with constant models in leaves. There is only a slight difference between the two (in favor of the former) and given that trees with linear models in the leaves are more difficult to interpret we only consider the ordinary regression tree results in the following.

Table 2: A regression tree for predicting chemical oxygen demand ($K_2Cr_2O_7$) from bioindicator data.

```
OLIGOCHAETA Lumbriculus variegatus <= 1 :
|   BACTERIA Sphaerotilus natans <= 3 :
|   |   DIPTERA Chironomus thummi > 1 : AV 29.64 (19)
|   |   DIPTERA Chironomus thummi <= 1 :
|   |   |   BACILLARIOPHYTA Nitzschia palea <= 0 :
|   |   |   |   OLIGOCHAETA Tubifex sp. > 1 : AV 11.72 (39)
|   |   |   |   OLIGOCHAETA Tubifex sp. <= 1 :
|   |   |   |   |   COLEOPTERA Elmis sp. > 0 : AV 4.49 (197)
|   |   |   |   |   COLEOPTERA Elmis sp. <= 0 :
|   |   |   |   |   |   HIRUDINEA Erpobdella octoculata > 0 : AV 10.21 (35)
|   |   |   |   |   |   HIRUDINEA Erpobdella octoculata <= 0 :
|   |   |   |   |   |   |   BACILLARIOPHYTA Diatoma hiemale v.mesodon > 1 : AV 2.86 (25)
|   |   |   |   |   |   |   BACILLARIOPHYTA Diatoma hiemale v.mesodon <= 1 :
|   |   |   |   |   |   |   |   BACTERIA Sphaerotilus natans > 0 : AV 9.40 (26)
|   |   |   |   |   |   |   |   BACTERIA Sphaerotilus natans <= 0 :
|   |   |   |   |   |   |   |   |   AMPHIPODA Gammarus fossarum <= 0 : AV 4.31 (68)
|   |   |   |   |   |   |   |   |   AMPHIPODA Gammarus fossarum > 0 : AV 6.91 (57)
|   |   |   BACILLARIOPHYTA Nitzschia palea > 0 :
|   |   |   |   PLECOPTERA Leuctra sp. <= 0 :
|   |   |   |   |   BACILLARIOPHYTA Nitzschia sigmoidea > 0 : AV 8.96 (84)
|   |   |   |   |   BACILLARIOPHYTA Nitzschia sigmoidea <= 0 :
|   |   |   |   |   |   BACTERIA Sphaerotilus natans <= 0 : AV 11.17 (125)
|   |   |   |   |   |   BACTERIA Sphaerotilus natans > 0 :
|   |   |   |   |   |   |   PISCES > 0 : AV 22.18 (20)
|   |   |   |   |   |   |   PISCES <= 0 :
|   |   |   |   |   |   |   |   BACILLARIOPHYTA Nitzschia palea <= 3 : AV 12.69 (96)
|   |   |   |   |   |   |   |   BACILLARIOPHYTA Nitzschia palea > 3 : AV 21.63 (17)
|   |   |   |   PLECOPTERA Leuctra sp. > 0 :
|   |   |   |   |   BACTERIA Sphaerotilus natans <= 1 : AV 7.05 (133)
|   |   |   |   |   BACTERIA Sphaerotilus natans > 1 : AV 13.21 (19)
|   BACTERIA Sphaerotilus natans > 3 :
|   |   AMPHIPODA Gammarus fossarum <= 0 : AV 38.24 (50)
|   |   AMPHIPODA Gammarus fossarum > 0 :
|   |   |   TRICHOPTERA Rhyacophila sp. <= 0 : AV 26.25 (26)
|   |   |   TRICHOPTERA Rhyacophila sp. > 0 : AV 10.02 (10)
OLIGOCHAETA Lumbriculus variegatus > 1 :
|   BACILLARIOPHYTA Navicula sp. <= 0 : AV 111.47 (6)
|   BACILLARIOPHYTA Navicula sp. > 0 : AV 8.60 (9)
```

Highest correlations (above 0.6) are achieved when predicting ammonia, biological oxygen demand, and chemical oxygen demand ($K_2Cr_2O_7$). The trees for $K_2Cr_2O_7$ and ammonia induced from the entire dataset are given in Tables 2 and 3.

Highest chemical oxygen demand (111.47 $\mu g/l$) is predicted if the taxon *Lumbriculus variegatus* from the genus Oligochaeta occurs frequently or abundantly and the taxon *Navicula sp.* does not occur in the sample. High oxygen demand (38.24 $\mu g/l$) is also predicted if *Lumbriculus variegatus* occurs at most incidentally, *Gammarus fossarum* does not occur, and the bacteria *Sphaerotilus natans* is abundant in the sample.

Highest ammonia concentrations are predicted when the taxon *Chironomus thummi* occurs frequently or abundantly: 13.35 $\mu g/l$ if the bacteria *Sphaerotilus natans* does not occur in the sample or 6.72 $\mu g/l$ if it does and the bacteria *Beggiatoa alba* occurs frequently or abundantly. Very high ammonia concentration (5.94 $\mu g/l$) is also predicted if *Chironomus thummi* does not occur or occurs incidentally, but *Sphaerotilus natans* occurs abundantly, while the taxa *Diatoma vulgare* and *Cymbella ventricosa* do not occur in the sample.

In general, the trees are in agreement with expert knowledge. The taxa *Lumbriculus variegatus*, *Chironomus thummi*, *Sphaerotilus natans*, and *Beggiatoa alba* are all used as indicators of heavily polluted waters (high chemical oxygen demand and high ammonia concentrations mean heavily polluted waters). The taxa *Gammarus fossarum*, *Navicula sp.*, *Cymbella ventricosa*, and *Diatoma vulgare*, on the other hand, are used as indicators of clean to moderately polluted waters.

Some expectations of the domain experts were, however, not fulfilled. For example, caddis flies do not appear in the ammonia tree, despite their tolerance of high ammonia concentrations. Also, *Tubifex tubifex* was expected to play a key role in the BOD tree. Note, however, that the taxon *Lumbriculus variegatus* of the same genus (Oligochaeta) plays such a role and that the taxon *Tubifex sp.* also appears in the tree. It is possible that not enough cases of *Tubifex tubifex* were identified to species level.

Table 3: A regression tree for predicting NH_4 concentration from bioindicator data.

```
DIPTERA Chironomus thummi <= 1 :
|   BACTERIA Sphaerotilus natans <= 3 :
|   |   HIRUDINEA Helobdella stagnalis <= 1 :
|   |   |   PLECOPTERA Leuctra sp. > 0 : AV 0.19 (350)
|   |   |   PLECOPTERA Leuctra sp. <= 0 :
|   |   |   |   CYANOPHYTA Oscillatoria sp. <= 1 :
|   |   |   |   |   OLIGOCHAETA Tubifex tubifex > 0 : AV 1.10 (17)
|   |   |   |   |   OLIGOCHAETA Tubifex tubifex <= 0 :
|   |   |   |   |   |   ISOPODA Asellus aquaticus > 1 : AV 0.57 (56)
|   |   |   |   |   |   ISOPODA Asellus aquaticus <= 1 :
|   |   |   |   |   |   |   OLIGOCHAETA Tubifex sp. > 3 : AV 0.84 (19)
|   |   |   |   |   |   |   OLIGOCHAETA Tubifex sp. <= 3 :
|   |   |   |   |   |   |   |   COLEOPTERA Elmis sp. <= 0 : AV 0.31 (316)
|   |   |   |   |   |   |   |   COLEOPTERA Elmis sp. > 0 : AV 0.18 (147)
|   |   |   |   CYANOPHYTA Oscillatoria sp. > 1 :
|   |   |   |   |   CHLOROPHYTA Stigeoclonium tenue <= 1 : AV 0.53 (22)
|   |   |   |   |   CHLOROPHYTA Stigeoclonium tenue > 1 : AV 2.85 (5)
|   |   HIRUDINEA Helobdella stagnalis > 1 :
|   |   |   DIPTERA Simulium sp. <= 0 : AV 3.59 (9)
|   |   |   DIPTERA Simulium sp. > 0 : AV 0.29 (10)
|   BACTERIA Sphaerotilus natans > 3 :
|   |   BACILLARIOPHYTA Diatoma vulgare > 0 : AV 0.82 (50)
|   |   BACILLARIOPHYTA Diatoma vulgare <= 0 :
o   |   |   BACILLARIOPHYTA Cymbella ventricosa <= 0 : AV 5.94 (13)
|   |   |   BACILLARIOPHYTA Cymbella ventricosa > 0 : AV 1.46 (8)
DIPTERA Chironomus thummi > 1 :
|   BACTERIA Sphaerotilus natans <= 0 : AV 13.35 (9)
|   BACTERIA Sphaerotilus natans > 0 :
|   |   BACTERIA Beggiatoa alba <= 1 : AV 2.55 (24)
|   |   BACTERIA Beggiatoa alba > 1 : AV 6.72 (6)
```

4 Discussion

We addressed the problem of inferring chemical parameters of river water quality from biological ones by using regression tree induction. Initial experiments indicate that ammonia concentration, biological oxygen demand and chemical oxygen demand can be predicted relatively successfully from bioindicator data. One should bear in mind that changes in the biota may be caused by short term fluctuations of chemical parameter values, meaning that it is impossible to completely determine the latter from the former. Nevertheless, our work is a step towards enabling selective chemical monitoring of river water quality.

A potential weakness of our work arises from the fact that the biological data are a mix of genera, family, and species data. Data at the family level typically take into account only samples where identification was not carried out down to species level. It is thus necessary to update the data at the family and genera levels to take into account data on species level taxa and then repeat the learning experiments.

Acknowledgements

The Hydrometeorological Institute of Slovenia provided the biological, physical and chemical data on Slovenian rivers used in this study. Thanks are due to Herbert A. Hawkes for comments on the regression trees.

References

[1] Cairns, J., Douglas, W.A., Busey, F., and Chaney, M.D. (1968). The sequential comparison index - a simplified method for non-biologists to estimate relative differences in biological diversities in stream pollution studies. *J. Wat. Pollut. Control Fed.*, 40: 1607–1613.

[2] De Pauw, N. and Hawkes, H.A. (1993). Biological monitoring of river water quality. In *Proc. Freshwater Europe Symposium on River Water Quality Monitoring and Control*, pages 87–111. Aston University, Birmingham.

[3] Džeroski, S., and Grbović, J. (1995). Knowledge discovery in a water quality database. In *Proc. First International Conference on Knowledge Discovery and Data Mining*, pages 81–86. AAAI Press, Menlo Park, CA.

[4] ISO-BMWP (1979). *Assessment of the Biological Quality of Rivers by a Macroinvertebrate Score.* ISO/TC147/SC5/WG6/N5, International Standards Organization.

[5] Kolkwitz, R. and Marsson, M. (1902). Grundsatze für die biologische Beurteilung des Wassers nach seiner Flora und Fauna. *Mitt. Prüfungsanst. Wasserversorg. Äbwasserein*, 1: 33–72.

[6] Quinlan, J.R. (1993) Combining instance-based and model-based learning. In *Proc. Tenth International Conference on Machine Learning*, pages 236–243. Morgan Kaufmann, San Mateo, CA.