Simultaneous prediction of multiple chemical parameters of river water quality with TILDE

Hendrik Blockeel[1], Sašo Džeroski[2] and Jasna Grbović[3]

[1] Katholieke Universiteit Leuven
Dept. of Computer Science
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
[2] Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
[3] Hydrometeorological Institute
Vojkova 1b, SI-1000 Ljubljana, Slovenia

Abstract. Environmental studies form an increasingly popular application domain for machine learning and data mining techniques. In this paper we consider some applications of decision tree learning in the domain of river water quality. More specifically, we study a) the simultaneous prediction of multiple physico-chemical properties of the water from its biological properties using a single decision tree (as opposed to learning a different tree for each different property – we call this approach predictive clustering) and b) the prediction of past physico-chemical properties of the river water from its current biological properties. We discuss some experimental results that we believe are interesting both to the application domain experts and to the machine learning community.

1 Introduction

The quality of surface waters, including rivers, depends on their physical, chemical and biological properties. The latter are reflected by the types and densities of living organisms present in the water. Based on the above properties, surface waters are classified into several quality classes which indicate the suitability of the water for different kinds of use (drinking, swimming, ...).

Although water quality is related to both biological and physico-chemical properties, it is well known that the physico-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 [6]. This has increased the relative importance of biological methods for monitoring water quality [7]. Many different methods for mapping biological data to discrete quality classes or continuous scales have been developed (for an overview, see [7]). 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. Examples are the Saprobic Index [14], which is used in many countries of Central Europe (e.g., Germany, Slovenia, ...), and the Biological Monitoring Working Party Score (BMWP) [13] and its derivative Average Score Per Taxon (ASPT), which are used in the United Kingdom.

The main problem with the biological indices described above is their subjectivity [18]. The computation of these indices makes use of weights and other numbers that were assigned to individual bioindicators by (committees of) expert biologists and ecologists and are based on the experts' knowledge about the ecological requirements of the bioindicator taxa, which is not always complete. The assigned bioindicator values are thus subjective and often inappropriate [19]. An additional layer of subjectivity is added by combining the scores of the individual bioindicators through ad-hoc procedures based on sums, averages, and weighted averages instead of using a sound method of combination. While a certain amount of subjectivity cannot be avoided (water quality itself is a subjective measure, as it is tuned towards the interests humans have in river water), this subjectivity should only appear at the target level (classification) and not at the intermediate levels described above. This may be achieved by gaining insight into the relationships between biological, physical and chemical properties of the water and its overall quality, which is currently a largely open research topic. To this aim data mining techniques can be employed [18, 11, 9].

We point out that the importance of gaining such insight stretches beyond water quality prediction. For instance, the problem of inferring chemical parameters from biological ones is practically relevant, especially in countries where extensive biological monitoring is conducted. Regular monitoring for a very wide range of chemical pollutants would be very expensive, if not impossible. On the other hand, biological samples may, for example, reflect an increase in pollution and indicate likely causes or sources of (chemical) pollution. The work described in this paper is situated at this more general level.

The remainder of the paper is organized as follows. Section 2 describes the goals of this study and the difference with earlier work. Section 3 describes the available data on the water quality of Slovenian rivers, as well as the experimental setup. Section 4 describes the machine learning tool that was used in these experiments. Section 5 presents in detail the experiments and their results and in Section 6 we conclude.

2 Goals of this study

In earlier work [10, 11] machine learning techniques have been applied to the task of inferring biological parameters from physico-chemical ones by learning rules that predict the presence of individual bioindicator taxa from the values of physico-chemical measurements, and to the task of inferring physico-chemical parameters from biological ones [9].

Džeroski *et al.* [9] discuss the construction of predictive models that allow prediction of a specific physico-chemical parameter from biological data. A different predictive model is built for each parameter. The models, which are constructed using Quinlan's M5 system [17], are in the form of regression trees. This approach is compared with nearest neighbour and linear regression methods; the authors conclude that the induction of regression trees is competitive with the other approaches as far as predictive accuracy is concerned, and moreover has the advantage of yielding interpretable theories.

A comparison of the different trees shows that the trees for different target variables are often similar, and that some of the taxa occur in many trees (i.e., they are sensitive to many physico-chemical properties). This raises the question whether it would be possible to predict many or all of the properties at once, with only one (relatively simple) tree, and without significant loss in predictive accuracy. As such, this application seems a good test case for recent research on simultaneous prediction of multiple variables [1].

A second extension with respect to the previous work is the prediction of past physico-chemical properties of the water; more specifically, the maximal, minimal and average values of these properties over a period of time. As mentioned before, physico-chemical properties of water give a very momentary view of the water quality; watching these properties over a longer period of time may alleviate this problem. This is the second scientific issue we investigate in this paper.

3 The Data

The data set we have used is the same one as used in [9]. The data come from the Hydrometeorological Institute of Slovenia (HMZ) that performs water quality monitoring for Slovenian rivers and maintains a database of water quality samples. The data provided by HMZ cover a six year period (1990–1995). Biological samples are taken *twice a year*, once in summer and once in winter, while physical and chemical samples are taken *several times a year* (periods between measurements varying from one to several months) for each sampling site.

The physical and chemical samples include the measured values of 16 different parameters: biological oxygen demand (BOD), electrical conductivity, chemical oxygen demand ($K_2Cr_2O_7$ and $KMnO_4$), concentrations of Cl, CO_2, NH_4, PO_4, SiO_2, NO_2, NO_3 and dissolved oxygen (O_2), alkalinity (pH), oxygen saturation, 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.

Our data are stored in a relational database represented in Prolog; in Prolog terminology each relation is a predicate and each tuple is a fact. The following predicates are relevant for this text:

- *chem(Site, Year, Month, Day, ListOf16Values)* : this predicate contains all physico-chemical measurements. It consists of 2580 facts.
- *bio(Site, Day, Month, Year, ListOfTaxa)*: this predicate lists the taxa that occur in a biological sample; ListOfTaxa is a list of couples (taxon, abundance-level) where the abundance level is 1, 3 or 5 (taxa that do not occur are simply left out of the list). This predicate contains 1106 facts.

Overall the data set is quite clean, but not perfectly so. 14 physico-chemical measurements have missing values; moreover, although biological measurements are usually taken on exactly the same day as some physico-chemical measurement, for 43 biological measurements no physico-chemical data for the same day are available. Since this data pollution is very limited, we have just disregarded the examples with missing values in our experiments. This leaves a total of 1060 water samples for which complete biological and physico-chemical information is available; our experiments are conducted on this set.

4 Predictive clustering and TILDE

Building a model for simultaneous prediction of many variables is strongly related to clustering. Indeed, clustering systems are often evaluated by measuring the average predictability of attributes, i.e., how well the attributes of an object can be predicted given that it belongs to a certain cluster (see, e.g., [12]). In our context, the predictive modelling can then be seen as clustering the training examples into clusters with small intra-cluster variance, where this variance is measured as the sum of the variances of the individual variables that are to be predicted, or equivalently: as the mean squared euclidean distance of the instances to their mean in the prediction space. More formally: given a cluster C consisting of n examples e_i that are each labelled with a target vector $\mathbf{x}_i \in I\!R^D$, the *intra-cluster variance* of C is defined as

$$\sigma_C^2 = 1/n \cdot \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}}) \quad (1)$$

where $\bar{\mathbf{x}} = 1/n \sum_{i=1}^{n} \mathbf{x}_i$.

In the above we assume the target vector to have only numerical components. This is not restrictive because nominal components can always be encoded as numbers (e.g. 0/1); for a nominal component with only two values minimising the variance corresponds to maximising the relative frequency of the most frequently occurring class, which is exactly what is done by most classification systems. (Note that most approaches to classification and regression are just special cases of predictive clustering, where $D = 1$ and the prediction space is nominal, respectively numerical. For rule-based systems, each rule body describes one cluster; for tree-based systems the leaves of the tree (in some approaches also the internal nodes) are clusters described by the tests in the tree.)

In our experiments we used the decision tree learner TILDE [2, 3]. TILDE is an ILP system[4] that induces so-called first order logical decision trees (FOLDT's).

[4] Inductive logic programming (ILP) is a subfield of machine learning where first order logic is used to represent data and hypotheses. First order logic is more expressive than the attribute value representations that are classically used by machine learning and data mining systems. From a relational database point of view, ILP corresponds to learning patterns that extend over multiple relations, whereas classical (propositional) methods can find only patterns that link values within the same tuple of a single relation to one another. We refer to [8] for details.

Such trees are the first-order equivalent of classical decision trees [2]. TILDE can induce classification trees, regression trees and clustering trees and can handle both attribute-value data and structural data. It uses the basic TDIDT algorithm [16], in its clustering or regression mode employing as heuristic the variance as described above. The system seemed fit for our experiments because of the following reasons:

- Most machine learning and data mining systems that induce predictive models can handle only single target variables (e.g., C4.5 [15], CART [5], M5 [17], ...). Building a predictive model for a multi-dimensional prediction space can be done using clustering systems, but most clustering systems consider clustering as a descriptive technique, where evaluation criteria are still slightly different from the ones we have here. (Using terminology from [12], descriptive systems try to maximise both predictiveness and predictability of attributes, whereas predictive systems maximise predictability of the attributes belonging to the prediction space.)
- Although the problem at hand is not, strictly speaking, an ILP problem (i.e., it can be transformed into attribute-value format; the number of different attributes would become large but not unmanageable for an attribute-value learner), the use of an ILP learner has several advantages:
  - No data preprocessing is needed: the data can be kept in their original, multi-relational format. This was especially advantageous for us because the experiments described here are part of a broader range of experiments, many of which would demand different and extensive preprocessing steps.
  - Prolog offers the same querying capabilities as relational databases, which allows for non-trivial inspection of the data (e.g., counting the number of times a biological measurement is accompanied by at least 3 physico-chemical measurements during the last 2 months, ...)

  The main disadvantage of ILP systems, as compared to attribute-value learners, is that they are less efficient; however, efficiency was not our prime concern here, and the inefficiency of ILP was not prohibitive and amply compensated for by the additional flexibility it offers.

5 Experiments

For all these experiments, TILDE was run with default parameters, except one parameter controlling the minimal number of instances in each leaf which was 20. From preliminary experiments this value was found to combine good predictive accuracy with reasonable tree size. All results reported here are obtained using 10-fold cross-validations.

5.1 Multi-valued Predictions

For this experiment we have run TILDE with two settings: predicting a single variable at a time (the results of which serve as a reference for the other setting),

	all variables (TILDE)	single variable (TILDE)	single variable (M5.1)
variable	r	r	r
T	0.482	0.563	0.561
pH	0.353	0.356	0.397
conduct.	0.538	0.464	0.539
O_2	0.513	0.523	0.484
O_2-sat.	0.459	0.460	0.424
CO_2	0.407	0.335	0.405
hardness	0.496	0.475	0.475
NO_2	0.330	0.417	0.373
NO_3	0.265	0.349	0.352
NH_4	0.500	0.489	0.664
PO_4	0.441	0.445	0.461
Cl	0.603	0.602	0.570
SiO_2	0.369	0.400	0.411
$KMnO_4$	0.509	0.435	0.546
$K_2Cr_2O_7$	0.561	0.514	0.602
BOD	0.640	0.605	0.652
avg	0.467	0.465	0.498

Table 1. Comparison of predictive quality of a single tree predicting all variables at once with that of a set of 16 different trees, each predicting one variable.

and predicting all variables simultaneously. When predicting all variables at once, the variables were first standardised ($z_x = (x - \mu_x)/\sigma_x$ with μ_x the mean and σ_x the standard devation); because standardised variables always have a variance of 1, this ensures that all target variables will be considered equally important for the prediction.[5] As a bonus the results are more interpretable for non-experts; e.g., "BOD=16.0" may not tell a non-expert much, but a standardised score of +1 always means "relatively high".

The predictive quality of the tree for each single variable is measured as the correlation of the predictions with the actual values. Table 1 shows these correlations; correlations previously obtained with M5.1 [9] are given as reference.

It is clear from the table that overall, the multi-prediction tree performs approximately as well as the set of 16 single trees. For a few variables there is a clear decrease in predictive performance (T, NO_2, NO_3), but surprisingly this effect is compensated for by the fact that some variables are predicted more accurately when they are predicted together with other variables (conductivity, CO_2, $KMnO_4$). A possible explanation for this is that when the variables to be predicted are not independent, they contain mutual information about one another

[5] Since the system minimises total variance, i.e. the sum of the variances of each single variable, the "weight" of a single variable is proportional to its variance; variables with small variance would not be considered important because reducing their variance would result in an insignificant reduction of the total variance.

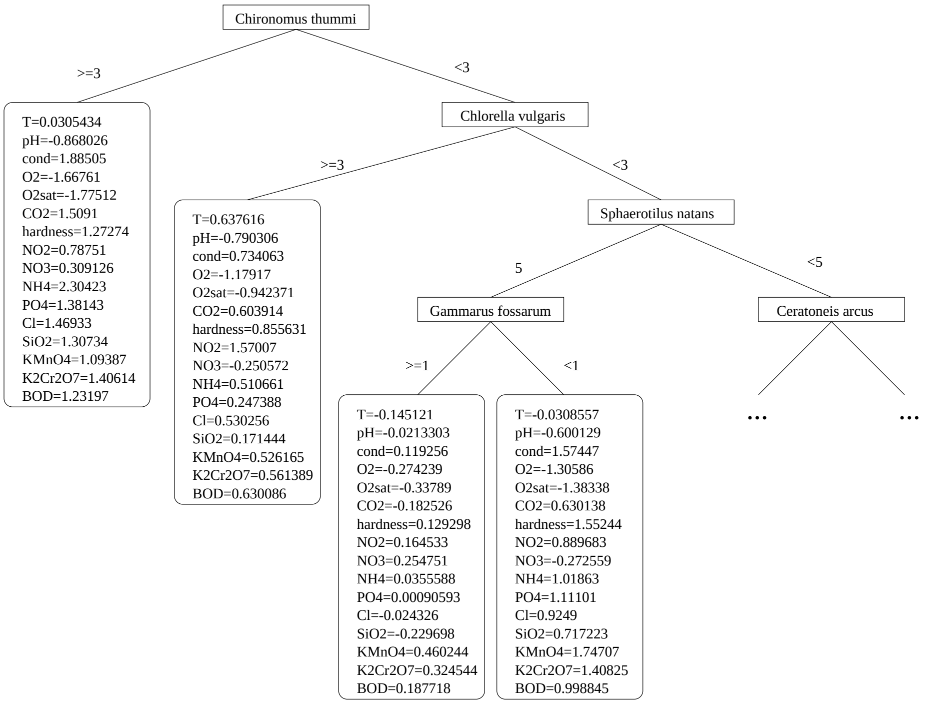

Fig. 1. An example of a clustering tree.

that may help the learner distinguish random fluctuations in a single variable from structural fluctuations. The table also shows that TILDE's performance is slightly worse than that of M5.1 (possibly because of different settings).

Note that because of the constant "minimal coverage" of 20, all trees have approximately equal size (about 35 nodes). This means that when predicting all 16 variables at once, the total theory size is effectively reduced by a factor of 16 when using the multi-prediction approach, with predictive accuracy suffering only very slightly from this.

Figure 1 shows the first levels of a multi-prediction tree that was induced during the experiment. The tree indicates, e.g., that *Chironomus thummi* has the greatest overall influence on the physico-chemical properties; its occurrence indicates low oxygen (saturation) levels, high conductivity, very high ammonia concentration, etc.

5.2 Predicting past values

In this experiment we try to predict the average, maximal and minimal values of physico-chemical parameters over a period of three months before the date when the biological sample was taken. Although three months is a relatively long

	available measurements			
months	2	3	4	5
2	536	90	6	0
3	672	311	77	2
4	759	444	147	21

Table 2. Overview of the number of SI measurements for which at least x physico-chemical measurements have been taken during the y months preceding the date of the SI measurement.

	minimum	maximum	average	current
variable	r	r	r	r
T	0.444	0.591	0.578	0.563
pH	0.351	0.316	0.355	0.356
conduct.	0.410	0.405	0.443	0.464
O_2	0.540	0.435	0.514	0.523
O_2-sat.	0.522	0.388	0.472	0.460
CO_2	0.359	0.401	0.403	0.335
hardness	0.412	0.451	0.497	0.475
NO_2	0.236	0.446	0.416	0.417
NO_3	0.313	0.359	0.336	0.349
NH_4	0.373	0.494	0.475	0.489
PO_4	0.271	0.400	0.418	0.445
Cl	0.513	0.311	0.413	0.602
SiO_2	0.344	0.432	0.394	0.400
$KMnO_4$	0.524	0.461	0.526	0.435
$K_2Cr_2O_7$	0.627	0.529	0.697	0.514
BOD	0.609	0.575	0.653	0.605
avg	0.428	0.437	0.474	0.465

Table 3. Comparison of predictive quality of trees when predicting the current value of a property vs. its minimal, maximal or average value during the last three months.

period (according to our domain expert 1 to 2 months would be optimal), for this data set we faced the problem that physico-chemical measurements are not always available for each month; in some cases the only measurement available for the last 5 months is taken on the same day as the biological measurement, which means that the minimal, maximal and average value over the period of time are equal to the current value. We quantify the problem in Table 2. This table shows an overview of the number of biological samples for which at least x physico-chemical measurements were available in the y months preceding the biological sample. By using a period of 3 months we ensure that for a reasonably-sized subset of the data set at least 2 or 3 measurements are available.

Results of this experiment are shown in Table 3. This table confirms most of the expert's expectations. For instance, for oxygen it was expected that the minimal oxygen level during a period of time, rather than its average or maximum, is most related to the biological data. Especially for O_2-saturation, and to a lesser extent for O_2, this is confirmed by the experiment. The expectation that for chemical oxygen demand ($KMnO_4$, $K_2Cr_2O_7$), the average value would be most important (because this parameter has a cumulative effect) is confirmed, although the minimal value also shows high correlation, which was not expected.

5.3 Discussion

Both experiments show the potential of decision tree learning for gaining insight in the water quality domain. The first experiment shows that simultaneous prediction of multiple parameters is feasible and increases the potential of decision trees for providing compact, interpretable theories. The second experiments confirms that it is possible to predict past properties of water from its current biological properties; moreover the results may lead to more insight into the mechanisms through which physico-chemical properties influence biological properties over a longer period of time.

6 Conclusions

We have used the decision tree learner TILDE to test two hypotheses: a) is it feasible to predict many properties at once with a single decision tree; b) is it feasible to predict past chemical properties from current biological data? In both cases the answer is positive. Our experiments globally confirm the expert's expectations, but here and there also contain some unexpected and interesting results. From the point of view of the water quality domain, some insight has been gained in the interdependencies of physico-chemical parameters and the way in which the properties of the water in the recent past can be predicted from current biological data. From the machine learning point of view, the feasability and potential advantages of a hitherto little explored technique, simultaneous prediction of multiple variables, has been demonstrated.

Related work in the machine learning domain includes the use of (descriptive) clustering systems for prediction of multiple variables [12]. In the application domain, we mention [9], [10] and [11] (on which this work builds further), and [4] which discusses a broad range of preliminary experiments in this domain.

There are many opportunities for further work: first of all some of the results described in this paper need to be studied in more detail by domain experts; secondly, simultaneous prediction of subsets of the 16 used variables, or of a mixture of current and past values, seems an interesting topic for further research; thirdly, many of the preliminary experiments described in [4], investigating other kinds of relationships in this domain, deserve further study.

Acknowledgements

The authors thank Damjan Demšar for his practical support and the Hydrometeorological Institute of Slovenia for making the data set available.

References

1. H. Blockeel. *Top-down induction of first order logical decision trees.* PhD thesis, Department of Computer Science, Katholieke Universiteit Leuven, 1998.
2. H. Blockeel and L. De Raedt. Top-down induction of first order logical decision trees. *Artificial Intelligence*, 101(1-2):285–297, June 1998.
3. H. Blockeel, L. De Raedt, and J. Ramon. Top-down induction of clustering trees. *Proc. 15th Int'l Conf. on Machine Learning*, pages 55–63, 1998.
4. H. Blockeel, S. Džeroski, and J. Grbović. Experiments with TILDE in the river water quality domain. Draft, Jožef Stefan Institute, Ljubljana, Slovenia, 1999.
5. L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, 1984.
6. J. Cairns, W.A. Douglas, F. Busey, and M.D. Chaney. 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, 1968.
7. N. De Pauw and H.A. Hawkes. Biological monitoring of river water quality. In *Proc. Freshwater Europe Symposium on River Water Quality Monitoring and Control*, pages 87–111. Aston University, Birmingham, 1993.
8. L. De Raedt. Attribute-value learning versus inductive logic programming: the missing links (extended abstract). *Proc. 8th Int'l Conf. on Inductive Logic Programming*, pages 1–8. Springer-Verlag, 1998.
9. S. Džeroski, D. Demšar, and J. Grbović. Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence*, 1999. In press.
10. S. Džeroski and J. Grbović. Knowledge discovery in a water quality database. *Proc. 1st Int'l Conf. on Knowledge Discovery and Data Mining (KDD'95).* AAAI Press, Menlo Park, CA, 1995.
11. S. Džeroski, J. Grbović, and W.J. Walley. Machine learning applications in biological classification of river water quality. In R.S. Michalski, I. Bratko, and M. Kubat, editors, *Machine Learning, Data Mining and Knowledge Discovery: Methods and Applications.* John Wiley and Sons, Chichester, 1997.
12. D. H. Fisher. Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research*, 4:147–179, 1996.
13. ISO-BMWP. Assessment of the biological quality of rivers by a macroinvertebrate score. Technical Report ISO/TC147/SC5/WG6/N5, International Standards Organization, 1979.
14. R. Pantle and H. Buck. Die biologische überwachtung der Gewas und die Darstellung der Ergebnisse. *Gas und Wasserfach*, 96:603, 1978.
15. J. Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann series in machine learning. Morgan Kaufmann, 1993.
16. J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
17. J.R. Quinlan. Combining instance-based and model-based learning. *Proc. 10th Int'l Workshop on Machine Learning.* Morgan Kaufmann, 1993.
18. W.J. Walley. Artificial intelligence in river water quality monitoring and control. In *Proc. Freshwater Europe Symposium on River Water Quality Monitoring and Control*, pages 179–193. Aston University, Birmingham, 1993.
19. W.J. Walley and H.A. Hawkes. A computer-based reappraisal of the biological moni toring working party scores using data from the 1990 river quality survey of engl and and wales. *Water Research*, 30:2086–2094, 1996.