ELSEVIER

Ecological Modelling 95 (1997) 95–111

ECOLOGICAL MODELLING

Using machine learning techniques in the construction of models. II. Data analysis with rule induction

Sašo Džeroski [a,*], Jasna Grbović [b], William J. Walley [c], Boris Kompare [d]

[a] *Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia*
[b] *Hydrometeorological Institute of Slovenia, Ljubljana, Slovenia*
[c] *School of Computing, Staffordshire University, Stafford, UK*
[d] *Institute of Sanitary Engineering, University of Ljubljana, Ljubljana, Slovenia*

Received 21 June 1995; accepted 18 March 1996

Abstract

Artificial intelligence and machine learning methods can be used to automate the acquisition of ecological knowledge, i.e., automate the construction of ecological models. This paper describes a particular methodology of machine learning, called rule induction, and its application to data analysis in several ecological domains. These include the biological classification of British rivers based on bioindicator data, the analysis of the influence of physical and chemical parameters on selected bioindicator organisms in Slovenian rivers and the biological classification of Slovenian rivers based on physical and chemical parameters as well as bioindicator data. In all three cases, valuable models (knowledge) in the form of rules were extracted from data acquired through environmental monitoring and/or expert interpretation of the acquired samples. This provides positive evidence for the utility of machine learning in ecological modelling.

Keywords: Artificial intelligence; Data analysis; Ecological modelling; Expert systems; Machine learning; Rule induction

1. Introduction

Artificial intelligence is the study of ways in which computers can perform tasks that demand intelligent behavior. Such tasks can be roughly divided in two categories: task that can be solved by commonsense reasoning and tasks that require expert knowledge. Example tasks from the first category are the understanding of stories or jokes written in natural language, recognition of objects on photographs, path finding in an environment with obstacles, and so on. The second category comprises tasks such as medical diagnostics, configuration of computer or telecommunication systems, diagnostics of different types of machinery, and so on. Ecological problems that belong to this category include the task of interpretation of biological or chemical samples of river water, i.e., classification of river water quality.

For solving problems from the second category, the methodology of expert systems was developed within the area of artificial intelligence. Expert systems are computer systems that can solve problems in a narrow domain with a performance similar to the one of human experts. The basic building blocks of an expert system are a knowledge base, an infer-

* Corresponding author. Tel.: +386-61-1773528; fax: +386-61-1251038; e-mail: saso.dzeroski@ijs.si.

PII S0304-3800(96)00029-4

ence engine, and a user interface (Harmon et al., 1988). The user interface takes care of the communication between the expert system and its user, including the acquisition of the data needed to solve a particular problem. The inference engine uses the acquired data and the rules in the knowledge base to infer intermediate and final conclusions regarding the solution of the problem.

The key part of the expert system is the knowledge base. While the other two parts of the expert system (called an expert system shell) are interchangeable across problem domains, the knowledge base is specific to a particular problem domain. It contains knowledge about that particular domain, typically formulated as a set of IF–THEN rules, which have an antecedent (IF) and a consequent (THEN) part. These rules can be viewed as a model of the domain studied, albeit in a form different from the traditional differential equation models. A knowledge base for an ecological domain can be regarded as an ecological model.

A model in the form of equations can be derived from expert knowledge and understanding of a particular domain. A knowledge base in the form of rules can be obtained in a similar manner. In the practice of building expert systems, this turned out to be a very difficult task. The construction of the knowledge base was mostly done through a process called knowledge acquisition (KA), during which a knowledge engineer (computer expert) tried to extract the knowledge from a domain expert (e.g., ecologist) by a permanent dialogue. This turned out to be a difficult and frustrating process. Namely, while experts perform their tasks with ease on particular problems in the given domain, they find it very difficult to describe how they do it precisely and in detail. And even when they gave a description, it often turned out that the description did not fit the practice, i.e., what the expert said he was doing was not what he was actually doing. This problem was so acute that it got a name of its own – the knowledge acquisition bottleneck.

A characteristic that defines a person as an expert in a particular area is his ability to solve problems in that area easily. Thus, examples of solved problems are not difficult to come by. Machine learning, i.e., learning from examples, can thus be a useful tool in constructing a knowledge base for an expert system, as indicated in a previous paper in this series (Kompare et al., 1994). From examples of solved problems (e.g., biological samples and the corresponding water quality classes), machine learning tools can construct general rules that capture the knowledge used to solve the problems. Such rules can constitute a knowledge base for an expert system. Expert systems of the second generation recognize the usefulness of machine learning (ML) and include ML tools, which can use examples of solved problems and additional knowledge about the domain, called deep knowledge.

Machine learning is a subarea of artificial intelligence that is concerned with improving the performance in a given domain based on existing experiences. It comprises a variety of different approaches, such as the induction of classification rules and trees, the induction of regression rules and trees, neural networks and machine discovery. The methods that are most often used in practice are the ones that induce classification and regression rules and trees.

A key issue in expert systems is the issue of transparency. An expert system should be able to explain its inferences and conclusions to the user. This is usually done by presenting the rules used in the reasoning involved in a particular problem and their interconnection. It is important to note that the rules have to be in symbolic and easily understandable format. While IF–THEN rules may be considered as such, neural networks are generally not transparent. This is the reason that we concentrate on rule induction.

The first paper of this series (Kompare et al., 1994) gives an introduction to the area of machine learning and illustrates the possibility to apply techniques from this area to ecological domains. This paper focuses on a particular type of machine learning techniques, called rule induction, and several applications of rule induction in ecological domains. Many other machine learning techniques, such as the induction of regression trees (Breiman et al., 1984) and machine discovery (Langley and Żytkow, 1989), can be applied to ecological domains (Kompare and Džeroski, 1994). Further papers in the series will describe these techniques and their applications to ecological domains.

This paper describes a system for rule induction called CN2 (Clark and Niblett, 1989, Clark and

Boswell, 1991, Džeroski et al., 1993). It then proceeds with a description of several ecological applications of this system. We are concerned in particular with problems related to river water quality.

The quality of surface waters, including rivers, depends on their physical, chemical and biological properties. The latter are reflected by the diversity and density of living organisms in the water. 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.

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. This has increased the importance of biological methods for monitoring water quality (De Pauw and Hawkes, 1993). Since Kolkwitz and Marsson (1902), who 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 (De Pauw and Hawkes, 1993; Grbović, 1994). Unfortunately, these methods have left much to be desired in terms of justification and reliability (Walley et al., 1992). This leaves room for the application of machine learning techniques with the goal to develop better methods of biological water quality classification.

We apply machine learning to several problems related to river water quality. These include biological classification of the quality of British and Slovenian rivers, as well as analyzing the influence of physical and chemical parameters on selected bioindicators. The latter task is addressed by inducing rules that predict the presence/absence of bioindicators from the measured values of physical and chemical parameters.

The paper is organized as follows. Section 2 gives a description of the task of rule induction, the rule induction system CN2, and the general experimental methodology used in our experiments. Section 3 describes the water quality data for British rivers, the experiments performed to induce classification rules and the results obtained. Section 4 first describes the data about Slovenian rivers, which include biological samples as well as samples of physical and chemical parameters. It then describes the different learning problems addressed and the results achieved. The problems of analyzing the influence of physical and chemical parameters on bioindicators in Slovenian rivers and biological water quality classification of Slovenian rivers are addressed. Section 5 briefly touches upon related work and Section 6 concludes with a comprehensive discussion.

2. Rule induction

2.1. The task

The task of rule induction is to find a set of IF–THEN classification rules from a given set of examples with known classifications. An example is described by the values of a fixed collection of features (attributes): A_i, $i \in \{1, \ldots, n\}$. Each attribute can either have a finite set of values (discrete) or take real numbers as values (continuous). A particular example e_j is thus a vector of attribute values: $e_j = (v_{1,j}, \ldots, v_{n,j})$. Each example is assigned one of N possible values of the class C (classifications): c_i, $i \in \{1, \ldots, N\}$. The class of example e_j will be denoted by c_j.

An IF–THEN rule has the form: IF condition THEN conclusion. The condition contains one or more attribute tests of the form $A_i = v_i$ for discrete attributes, respectively $A_i < v_i$ or $A_i > v_i$ for continuous attributes. The conclusion part has the form $C = c_i$. We say that an example is covered by a rule if the attribute values of the example obey the conditions in the IF part of the rule. The example is a positive example if its class agrees with the class predicted by the rule and a negative example otherwise.

For illustration, consider the rule-induction problem defined on the 1984 US Congressional Voting Records Database (Džeroski et al., 1993). This database contains the votes for each of the U.S. House of Representatives congressmen on the 16 key votes in 1984. Given the votes (attributes), the task is to predict the party affiliation (class) of a congressman, which may be either republican or democrat. Thus there are two classes. As each vote can be either for or against, the attributes are discrete (binary).

```
IF      adoption_of_the_budget_
        resolution=y
    AND physician_fee_freeze=n
    AND education_spending=n
THEN    class=democrat
```

The above rule was derived by the CN2 rule-induction system (Clark and Boswell, 1991) from the examples in the Congressional Voting Records Database. It predicts that a congressman is a democrat if he voted for the adoption_of_the_budget_resolution, against the physician_fee_freeze and against education_spending. The rule covers almost two hundred positive examples, i.e., correctly classifies almost two hundred congressmen from the database as democrats.

2.2. The CN2 algorithm

The CN2 system (Clark and Niblett, 1989; Clark and Boswell, 1991) uses the covering approach to construct a set of rules for each possible class: it constructs a rule that correctly classifies some examples, removes the correctly classified examples from the training set and repeats the process until no more examples remain. To construct a rule that classifies examples in a given class, CN2 starts with a rule with an empty antecedent (IF part) and the selected class as a consequent (THEN part). The antecedent of this rule is satisfied by all examples in the training set, and not only those of the selected class. CN2 then progressively refines the antecedent by adding conditions to it, until only examples of the selected class satisfy the antecedent. To allow for handling imperfect data, CN2 may construct a set of rules which is imprecise, i.e., does not classify all examples in the training set correctly.

Consider a partially built rule. The conclusion part is already fixed and there are some (possibly none) conditions in the IF part. The examples covered by this rule form the current training set. For discrete attributes, all conditions of the form $A_i = v_i$, where v_i is a possible value for A_i, are considered for inclusion in the condition part. For continuous attributes, all conditions of the form $A_i < (v_{ik} + v_{i(k+1)})/2$ and $A_i > (v_{ik} + v_{i(k+1)})/2$ are considered, where v_{ik} and $v_{i(k+1)}$ are two consecutive values of attribute A_i that actually appear in the current training set. For example, if the values 4.0, 1.0 and 2.0 for attribute A appear in the current training set, the conditions $A < 1.5$, $A > 1.5$, $A < 3.0$ and $A > 3.0$ will be considered.

Note that both the structure (set of attributes to be included) and the parameters (values of the attributes for discrete ones and boundaries for the continuous ones) of the rule are determined by CN2. Which condition will be included in the partially built rule depends on the number of examples of each class covered by the refined rule and the heuristic estimate of the quality of the rule. The heuristic estimates are mainly designed to estimate the performance of the rule on unseen examples in terms of classification accuracy. This is in accord with the task of achieving high classification accuracy on unseen cases.

Suppose a rule covers p positive and n negative examples. Its accuracy can be estimated by relative frequency as $p/(p+n)$. This heuristic is used in AQ15 (Michalski et al., 1986). It prefers rules which cover examples of only one class. The problem with this metric is that it tends to select very specific rules supported by only a few examples. In the extreme case, a maximally specific rule will cover (be supported by) one example and hence have an unbeatable score using the metrics of apparent accuracy (scores 100% accuracy). Apparent accuracy on the training data, however, does not adequately reflect true predictive accuracy, i.e., accuracy on new testing data. It has been shown (Holte et al., 1989) that rules supported by few examples have very high error rates on new testing data.

The problem lies in the estimation of the probabilities involved, i.e., the probability that a new example is correctly classified by a given rule. If we use relative frequency, the estimate is only good if the rule covers many examples. In practice, however, not enough examples are available to estimate these probabilities reliably at each step. Therefore, probability estimates that are more reliable when few examples are given should be used.

The newer version of CN2 (Clark and Boswell, 1991) uses the Laplace estimate to estimate the accuracy of rules. This estimate is more reliable than relative frequency. If a rule covers p positive and n negative examples its accuracy is estimated as $(p+1)/(p+n+N)$, where N is the number of possible classes.

Unfortunately, the Laplace estimate relies on the assumption that all classes are equally probable a priori, an assumption which is rarely true in practice. We have therefore extended CN2 to enable the use of an even more sophisticated probability estimate, named the *m*-estimate (Džeroski et al., 1993). The *m*-estimate takes into account the prior probabilities of each class, and combines them with the evidence provided by the examples covered by the particular rule. The parameter m controls the role of the prior probabilities and the evidence provided by the examples: higher m gives more weight to the prior probabilities and less to the examples. In a sense, a higher m means that we trust the examples less, i.e., we consider them to be 'noisier'. If a rule that predicts class c_i covers p positive and n negative examples, its accuracy is estimated to be $(p + mp_i)/(p + n + m)$ (Cestnik, 1990), where p_i is the prior probability of class c_i. In CN2, p_i is estimated from the complete training set by relative frequency.

CN2 can also use a significance measure to enforce the induction of reliable rules. A rule is deemed reliable (significant) if the class distribution of the examples it covers is significantly different from the prior class distribution as given by the entire training set. This is measured by the likelihood ratio statistic (Clark and Boswell, 1991). Suppose the rule covers r_i examples of class c_i, $i \in \{1, \ldots, N\}$. Let $q_i = r_i/(r_1 + \ldots + r_N)$ and let p_i be the prior probability of class c_i. The value of the likelihood ratio statistic is then

$$2(r_1 + \ldots + r_N) \sum_{i=1}^{N} q_i \log_2(q_i/p_i).$$

This statistic is distributed as χ^2 with $N-1$ degrees of freedom. If its value is above a specified significance threshold, the rule is deemed significant.

CN2 can induce a set of IF–THEN rules which is either ordered or unordered. In the first case, the rules are checked precisely in the order specified during classification. Given an example to classify, the class predicted by the first rule that covers the example is returned. In the second case, all rules are checked and all the rules that cover the example are taken into account. Conflicting decisions are resolved by taking into account the number of examples of each class (from the training set) covered by each rule. Suppose we have a two class problem and two rules with coverage [10, 2] and [4, 40] apply, i.e., the first rule covers 10 examples of class c_1 and 2 examples of class c_2, while the second covers 4 examples of class c_1 and 40 examples of class c_2. The 'summed' coverage would be [14, 42] and the example is assigned class c_2. Our version of CN2 (Džeroski et al., 1993) can also return a probability distribution as an answer: this is the most general answer a classifier can return. For our hypothetical example, we divide the coverage [14, 42] with the total number of examples covered (56) and obtain as answer the probability distribution [0.25, 0.75]. This means that the probability of the example belonging to class c_1 is $1/4$, while for c_2 that probability is $3/4$.

Another feature of our extended version of CN2 is the possibility to measure the information score (Kononenko and Bratko, 1991) of induced rules. The information score is a performance measure which is not biased by the prior class distribution. It accounts for the possibility to achieve high accuracy easily in domains with a very likely majority class: classifying into the majority class all the time gives a zero information score.

Let the correct class of example e_k be c, its prior probability p and the probability returned by the classifier p'. The information score of this answer is

$$I(e_k) = \begin{cases} -\log p + \log p' & p' \geq p \\ \log(1-p) - \log(1-p') & p' < p \end{cases}$$

As $I(e_k)$ indicates the amount of information about the correct classification of e_k gained by the classifier's answer, it is positive if $p' > p$, negative if the answer is misleading $p' < p$ and zero if $p' = p$.

The *relative information score* I_r of the answers of a classifier on a testing set consisting of examples $e_1, e_2, \ldots, e_t$ belonging to one of classes $c_1, c_2, \ldots, c_N$ can be calculated as the ratio of the *average information score of the answers* and the *entropy of the prior distribution of classes.*

$$I_r = \frac{\frac{1}{t} \times \sum_{k=1}^{t} I(e_k)}{-\sum_{i=1} pi \log pi)}$$

The relative information score of the 'default' classifier, which always returns the prior probability distri-

bution, is zero. If the classifier always guesses the right class and is absolutely sure about it $p' = 1$, then $I_r = 1$, provided the class distributions of the training and testing sets are the same. In the remainder of the paper we will use the terms information score and information content with the meaning relative information score.

2.3. The experimental methodology

In the experiments described in this paper, CN2 was used to induce sets of unordered rules. The rules were required to be highly significant (at the 99% level) and thus reliable. Except for the significance threshold and the search heuristic settings, described below, the parameter settings of CN2 were the default ones (see Clark and Boswell, 1991).

To select the appropriate probability estimation method, i.e., the appropriate value for the parameter m, we used the following methodology. Fifteen different values of the parameter m were tried (0, 0.01, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024), as suggested by earlier experiments (Cestnik, 1990; Džeroski et al., 1993). The Laplace estimate was also tried out. For a given set of examples, we thus induced 16 sets of rules and chose the best according to the following lexicographic criterion: (1) information score, (2) accuracy, (3) smaller value of the parameter m. The accuracy and the relative information score are estimated on the training set.

This procedure allows us to choose the right level of fitting: overfitting is prevented by applying the significance threshold. Preliminary experiments showed that as the parameter m increases, the accuracy and information score of the induced rules increase until an optimum is reached; further increasing m causes a decrease in the accuracy and information score (Ličan-Milošević, 1994).

Note that a behavior of this kind is obtained only if we use a high significance threshold. If we do not apply a significance threshold then accuracy and information score fall as m grows: this prevents us from being able to choose an appropriate value of m on the training set. Earlier experiments chose an appropriate value for m on the testing set (Cestnik, 1990; Džeroski et al., 1993), which is a methodological flaw.

3. Biological classification of British rivers

To arrive at a suitable method for biological water quality classification, one could use the following approach. Suppose we had a set of samples (that list the present bioindicators and their abundance) and their correct classifications. We could then use a rule induction method to capture the knowledge needed to classify samples correctly. Correct classification may be provided by an expert in riverine ecology.

A similar approach has been used by Walley et al. (1992) and Ruck et al. (1993). They used field samples of benthic communities taken from the upper Trent catchment in England, classified into water quality terms by an expert river ecologist. Walley et al. (1992) used statistical methods (Bayesian classifiers) to reproduce the expert behavior, while Ruck et al. (1993) trained a neural network to do the same. Unfortunately, neither the Bayesian classifier, nor the neural networks are easy to understand and interpret. We therefore decided to use rule induction on the same problem to produce symbolic and transparent rules.

3.1. The data

The data considered here consist of 292 samples of benthic macroinvertebrates collected as part of a biological monitoring programme of the British National River Authority. They originate from the upper Trent catchment in England. They are given in the form of a site by species matrix, where the rows correspond to samples (sites) and the columns correspond to eighty different macroinvertebrate families (or taxa, in some cases). For each sample, the abundances are given for each of the eighty families of invertebrates.

The abundance of animals found is recorded as an integer between 0 to 6. Zero denotes that no members of the particular family were found in the sample, 1 denotes the presence of 1–2 members, 2 denotes 3–9 members present, 3 denotes 10–49, 4 denotes 50–99, 5 denotes 100–999, and 6 denotes more than 1000. There is a large number of zeros in the matrix (it is quite sparse), as most of the families are absent from any given sample. In our experiments, the abundances of all families were treated as continuous variables.

The samples were classified by the riverine ecology expert H.A. Hawkes, who is now an Honorary Reader at Aston University, Birmingham, UK. He was a member of the Biological Monitoring Working Party (De Pauw and Hawkes, 1993), the committee which was set up by the Department of Environment to establish a biological monitoring system for the UK. He chaired the working sub-committee responsible for the allocation of the family scores, intended to reflect the relative importance and influence of individual families on the overall water quality. In other words, he is as an acknowledged expert in the area of biological classification of British rivers.

The samples were classified into five classes, based on the level of organic pollution indicated by the invertebrate community. This was originally done as part of a project to develop a Bayesian-based expert system (Walley et al., 1992). The five classes were designed to mirror the five chemical classes (1a, 1b, 2, 3, and 4) presently in use in the UK, and were designated B1a, B1b, B2, B3, and B4 to distinguish them from the chemical classes (Ruck et al., 1993). Class B1a indicates least polluted (best quality) water, while class B4 represents water of the poorest quality. Table 1 shows two biological samples and their classifications as assigned by the expert. In machine learning terminology, the abundance of each family is an attribute, each biological sample is an example, and the classification assigned by the expert is the class.

3.2. The experiments

We addressed two learning problems: predicting the water quality class from the original eighty attributes and predicting the water quality class in the presence of additional attributes that describe the diversity of the population in a given sample. A set of rules was generated for each of the problems using the methodology described above. The rules were then inspected and evaluated by a domain expert (H.A. Hawkes). Their performance was also measured in terms of classification accuracy and information content.

CN2 (Džeroski et al., 1993) induced 12 rules from the 292 samples containing 80 attribute values each. The best value of the m-parameter turned out to be $m = 32$. As indicated in Section 2, a significance threshold of 99% was employed, enforcing the induction of reliable rules. On the average, each rule covered twenty-five examples and contained five conditions. These rules achieve 86.3% accuracy on the training set, as well as 75% information content. Three of the 12 rules are shown in Tables 2–4.

The induced rules mention the presence/absence or give other bounds on the abundance levels for

Table 1
Two biological samples from the upper Trent catchment in England. The expert classified Sample 74698 into class B1a and Sample 75892 into class B1b. (a) The complete samples represented as lists of (family, abundance) pairs. (b) An excerpt from the site by species matrix representation of the two samples

(a)

Sample 74698 (class B1a)
(Planariidae, 3), (Ancylidae, 2), (Hydracarina, 3), (Gammaridae, 2), (Baetidae, 2), (Heptageniidae, 3), (Leptophlebiidae, 2), (Nemouridae, 3), (Leuctridae, 2), (Perlodidae, 1), (Chloroperlidae, 1), (Rhyacophilidae, 1), (Hydropsychidae, 2), (Limnephilidae, 3), (Chironomidae, 1)
Sample 75892 (class B1b)
(Hydrobiidae, 4), (Physidae, 3), (Lymnaeidae, 1), (Ancylidae, 2), (Sphaeriidae, 3), (Glossiphoniidae, 2), (Erpobdellidae, 2), (Hydracarina, 3), (Gammaridae, 1), (Hydropsychidae, 4), (Tipulidae, 2)

(b)

SampleID	Planariidae ...	Hydracarina	Gammaridae ...	Tipulidae ...	Class
74698	3	3	2	0	B1a
75892	0	3	1	2	B1b

Table 2
A rule that predicts the water quality class B1a. It covers 42 examples of class B1a and no examples of the other four classes, as indicated by the numbers between the square brackets. The condition (IF) part states that less than 50 animals (≤ 3) of the family Hydrobiidae are present in the sample, Planorbidae are absent (≤ 0), there are less than 1000 animals of the family Gammaridae (≤ 5), and there are some Leuctridae (> 0) present

IF	Hydrobiidae $\leq$ 3
AND	Planorbidae $\leq$ 0
AND	Gammaridae $\leq$ 5
AND	Leuctridae > 0
THEN	Class = B1a
	[42 0 0 0 0]

thirty-five out of the eighty families. The rule induction algorithm chooses the families that are most indicative of the water quality class, mainly in conjunction with other families. This is in contrast with standard practices, which usually consider each bioindicator independently and then combine the evidence of different bioindicators through different kinds of indices (De Pauw and Hawkes, 1993).

To test the consistency of the induced rules with the existing expert knowledge, the twelve rules were presented to the ecological expert without the conclusion part. The order in which the rules were presented was random. The expert was then required to specify the appropriate classes for the conclusions of the rules. Most of the time, the expert's conclusions confirmed the rules: for five rules he gave the same class, for three he specified the next worse class, for three he specified a possible range of the correct class and the next worse, and for one rule one class better than the actual. For the rule given in Table 2, the expert specified class B1b (actual B1a),

Table 3
A rule that predicts the water quality class B2. It covers 41 examples of class B2. It requires the absence of Scirtidae, at least 50 members of Asellidae, and the presence of Gammaridae, which at the same time must not be overabundant (less than 100 members)

IF	Asellidae > 2
AND	0 < Gammaridae $\leq$ 4
AND	Scirtidae $\leq$ 0
THEN	Class = B2
	[0 0 41 0 0]

Table 4
A rule that predicts the water quality class B3. Note that this rule covers ten examples of class B4 and three of class B2, in addition to the twenty-eight examples of class B3. This fact is taken into account during the classification process, which combines all rules that correspond to the example classified. The rule relies heavily on the absence of several families (Planariidae, Lumbricidae, Gammaridae, Veliidae, Hydropsychidae, Simulidae, and Muscidae). It requires the presence of Tubificidae and Asellidae, and restricts the number of Glossiphoniidae to be less than 10

IF	Planariidae $\leq$ 0
AND	Tubificidae > 0
AND	Lumbricidae $\leq$ 0
AND	Glossiphoniidae $\leq$ 2
AND	Asellidae > 0
AND	Gammaridae $\leq$ 0
AND	Veliidae $\leq$ 0
AND	Hydropsychidae $\leq$ 0
AND	Simulidae $\leq$ 0
AND	Muscidae $\leq$ 0
THEN	Class=B3
	[0 0 3 28 10]

for the rule in Table 3 class B3 (actual B2), and for the rule in Table 4 class B3 (actual B3).

While the rules were roughly consistent with expert knowledge, some criticism was expressed regarding the reliance of the rules on the absence of families from samples (see for example the rule in Table 4). The absence of a taxon may often be insignificant and in many cases, but not all, provides little extra information. The main criticism, however, was that the rules use only a small number of taxa, whereas the expert takes into account the whole community when giving his classification. The diversity of the community structure is an important indication of the water quality. The expert was reluctant to interpret some of the rules because he felt that he needed more information in order to draw a proper conclusion.

Taking into account the criticism, we conducted an experiment where the learning system was given six additional attributes intended to capture the diversity of a sample. The attributes, named (MoreThan0, ..., MoreThan5), reflect the number of families present over a certain abundance level: MoreThan0 is the total number of families present, while MoreThan5 is the number of families present with at least 1000 members in the sam-

ple. The same settings as above were used, except that the best value for the parameter m was $m = 64$. Thirteen rules were generated with accuracy 88.4% (on the training set) and information content 80%. This performance improvement suggests that the expert criticism was justified. An example rule that exploits the additional diversity attributes is given in Table 5.

To estimate the performance on unseen cases, we split the set of 292 examples randomly into a training set of 195 examples (two thirds) and a testing set of 97 examples (one third). A set of rules was generated from the training set using the standard methodology and then evaluated on the testing set. For the original learning problem (80 attributes), the value of $m = 128$ turned out to be the best. Ten rules were induced with an accuracy of 61.9% and information score of 55% on the unseen cases. For comparison, the performance on the training set was 91.3% in terms of accuracy and 80% in terms of information content. For the extended learning problem (86 attributes), $m = 32$ was the best setting. An accuracy of 68.0% and an information score of 56% were achieved on unseen cases (93.8% and 83%, respectively, on the training set) by the 12 induced rules. There is an obvious performance improvement from the original to the extended problem, especially in terms of classification accuracy on unseen cases.

For comparison to more classical classification methods, we applied the nearest neighbour algorithm to our two classification problems. The nearest neighbour (NN) algorithm is one of the most well-known classification algorithms and an enormous body of research exists on the subject, as evidenced by the survey of Dasarathy (1990). In essence, the NN algorithm treats attributes as dimensions of an Euclidean space and examples as points in this space. In the training phase, the classified examples are stored without any processing. When classifying a new example, the Euclidean distance between that example and all training examples is calculated and the class of the closest training example is assigned to the new example.

The more general kNN method takes the k nearest training examples and determines the class of the new example by majority vote. In improved versions of kNN, the votes of each of the k nearest neighbours are weighted by the respective proximity to the new example. An optimal value of k may be determined automatically from the training set by using leave-one-out cross-validation. Finally, the contribution of each attribute to the distance may be weighted by the mutual information between that attribute and the class. These so-called feature weights can also be determined automatically on the training set. For our experiments, we used the kNN algorithm as implemented by Wettschereck (1994), which includes the improvements described above.

Using the same 195 training examples and 97 testing examples as CN2, different versions of kNN performed as follows. The basic NN algorithm achieved 55.7% accuracy without and 59.8% with diversity attributes. The kNN method without feature weights achieved 55.7% accuracy in both cases, the best value of k being 3. Finally, the kNN method with feature weights achieved 56.7% and 58.8% accuracy, respectively, the best values of k being 3 and 5, respectively. The rules induced by CN2 obviously perform better in both cases, which means that they represent useful generalizations of the training examples.

Table 5
A rule that uses the diversity attributes. It predicts class B4 (poorest quality). At most five different taxa are allowed in the sample (MoreThan0 ≤ 5), and at least two of these have to be present with at least three animals (MoreThan1 > 1). Dixidae have to be absent, Asellidae may be present with up to 50 and Oligochaeta with up to 1000 animals. It covers 25 examples of class B4 and four examples of class B3

IF	Oligochaeta ≤ 5
AND	Asellidae ≤ 3
AND	Dixidae ≤ 0
AND	MoreThan0 ≤ 5
AND	MoreThan1 > 1
THEN	Class = B4
	[0 0 0 4 25]

4. Slovenian rivers

According to current legislation on water classification in Slovenia (Official, 1978), the water belongs to the first (best) quality class if it is suitable for drinking, bathing and fisheries. Second class water is

suitable for fisheries and recreation, including bathing; after simple treatment (coagulation, filtration, disinfection) it can be used for industrial purposes, even in the food industry. Third class waters can be used for irrigation and (after conditioning) in the industry, except the food industry. Water of the fourth (worst) quality class can be used only for purposes less demanding than the above ones and after appropriate treatment.

Slovenian water authorities use the saprobic index method, as introduced by Pantle and Buck (1955) and modified by Zelinka and Marvan (1961), to map biological data to a continuous scale. The saprobic index derived from a given water sample is thus a single real number between one and four that reflects the quality of the water. For presentation purposes, the saprobic index is mapped into a discrete quality scale of four basic classes and three intermediate classes, i.e., seven discrete classes: 1., 1.–2., 2., 2.–3., 3., 3.–4., and 4. Class 1. corresponds to clean waters where the saprobic index ranges between 1.00 and 1.50, while class 4. corresponds to heavily polluted waters where the saprobic index ranges between 3.51 and 4.00. The four basic classes correspond to the legislation defined classes, but are somewhat different, as the latter rely mainly on chemical properties.

The saprobic index is calculated as a weighted average of the densities of a selected set of living organism families (or other taxonomical units, referred to as taxa). The taxa used are such that their biology, importance and ecological role is known. Such taxa are called bioindicators, since they reflect the overall water quality as affected by physical and chemical influences over a period of time. The ecological role and water quality importance is not known for many taxa and may furthermore differ from country to country (Grbović, 1994). Little is also known about the influence of physical and chemical water properties on many taxa. From an ecological and water quality point of view, these are important research topics.

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 four year period, from 1990 to 1993. 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 fifty different parameters, which include for example dissolved oxygen and hardness, while the biological samples include a list of all taxa present at the sampling site and their density. The density (abundance level) 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. Biological samples include the corresponding saprobic index value and the corresponding quality class as determined by the index. In total, 698 water samples were available on which both physical/chemical and biological analyses were performed: our experiments were conducted using these samples.

Given the data described above, we used the CN2 rule induction system to search for new knowledge. We formulated several learning problems: analysis of the influence of selected physical and chemical water properties on the presence of selected taxa; water quality classification starting from a selected set of bioindicators; and water quality classification based on a selected set of physical and chemical properties. In Sections 4.1 and 4.2, we briefly describe the methodology used to learn and evaluate rules and then present the results for each of the learning problems. The evaluation of induced rules comprises comments by Jasna Grbović, an expert biologist that performs analyses of biological samples at HMZ and has rich knowledge on the ecology of plants and animals found in Slovenian rivers, as well as the classification accuracy and information score of the rules (estimated on unseen cases).

For each of the learning problems, two sets of experiments were performed. The first set induced rules from all 698 examples, aiming to find as much reliable patterns (and hopefully knowledge) as possible. The rules derived in this way were inspected by the expert biologist and evaluated in the light of existing knowledge on riverine ecology and water quality. The second batch of experiments was aimed at evaluating the performance of induced rules in terms of their accuracy and information score on

unseen cases. To this end, we split the entire dataset into a training (70% – 489 cases) and testing (30% – 209 cases) set in ten different ways. For each split, we induced a set of rules according to the above methodology from the training set, then tested the performance of the rules on the test set; the results appearing in Sections 4.1 and 4.2 are the averages over the ten different splits.

4.1. The influence of physical and chemical parameters on selected organism

Plants are more or less influenced by the following physical and chemical parameters (water properties): total hardness, nitrogen compounds (NO_2, NO_3, NH_4), phosphorus compounds (PO_4), silica (SiO_2), iron (Fe), surfactants (detergents), chemical oxygen demand (COD), and biochemical oxygen demand (BOD). The last two parameters indicate the degree of organic pollution: the first reflects the total amount of degradable organic matter, while the second reflects the amount of biologically degradable matter. Animals are mostly influenced by a different set of parameters: water temperature, acidity or alkalinity (pH), dissolved oxygen (O_2, saturation of O_2), total hardness, chemical (COD), and biochemical oxygen demand (BOD).

The experiments presented in this section studied the influence of the physical and chemical parameters listed above on ten plant taxa and seven animal taxa. On the plant side, eight kinds of diatomae (BACILLARIOPHYTA) and two kinds of green algae (CHLOROPHYTA) were studied. The animal taxa chosen for study include worms (OLIGOCHAETA), crustacea (AMPHIPODA) and five kinds of insects.

For each of the selected taxa we defined an attribute-based learning problem, the attributes being the selected physical and chemical parameters (`Hardness, NO2, NO3, NH4, PO4, SiO2, Fe, Detergents, COD, BOD` for plants; `Temperature, PH, O2, Saturation, COD, BOD` for animals). The class is the presence of the selected taxon (with values `Present` and `Absent`). Seventeen different learning problems (domains) were thus defined.

We now summarize the experiments with the above learning problems, carried in accordance with the methodology specified above. We first give an overview of the performance of the induced rules, both for rules derived from the whole data set (calculated on the training set) and for rules derived from the ten splits (calculated on the testing set and averaged over the ten splits). We then give excerpts of the expert rule evaluation for selected plant and animal taxa. A detailed description of the selected taxa, the experiments, the performance of induced rules, the selected sets of rules, and the expert evaluation of these rules can be found in the BSc Thesis of Ličan-Milošević (1994).

Table 6 gives an overview of the performance of the rules as evaluated on the whole dataset (W) and the ten splits (P) into a training and testing set. The accuracy on the whole (training) dataset ranges between 66 and 85% (the default accuracy, i.e., the

Table 6
The performance of rules predicting taxa presence from physical and chemical parameters. The accuracy of the induced rules (A), as well as their information score (I), expressed as percentages, are given. The frequency of the majority class (D) in the training set is also given for comparison. Experiments on all available data were performed (W), where accuracies and information scores are measured on the training set. For each taxon, ten experiments were performed with 30% of the data withheld from the training phase and performance evaluated on these 30% of unseen cases. Averages of the ten experiments are given (P)

Taxon	D		A		I	
	W	P	W	P	W	P
Plant taxa						
Coconeis placentula	55.9	58.4	77.8	65.5	43	20.0
Cymbella sp.	65.8	64.6	77.7	67.7	33	16.6
Cymbella ventricosa	56.3	59.3	75.6	63.0	38	15.1
Diatoma vulgare	59.7	56.5	79.2	64.6	38	14.6
Navicula cryptocephala	63.2	61.7	78.4	62.6	41	12.1
Navicula gracilis	66.6	71.8	79.8	71.3	41	16.9
Nitzschia palea	60.2	60.8	78.8	69.7	50	29.2
Synedra ulna	50.0	48.3	69.1	53.1	43	7.0
Cladophora sp.	54.6	59.8	66.3	56.8	34	11.4
Oedogonium sp.	69.5	68.4	80.8	67.6	35	5.1
Animal taxa						
Tubifex sp.	67.2	66.0	85.1	70.0	49	20.5
Gammarus fossarum	56.7	55.0	75.9	64.6	35	17.6
Baetis sp.	65.9	67.9	77.9	65.7	31	7.0
Leuctra sp.	68.2	68.9	85.1	70.6	47	17.9
Chironomidae (green)	55.3	50.2	67.2	53.3	23	1.4
Simulium sp.	65.2	75.6	75.5	65.4	28	4.6
Elmis sp.	65.8	64.6	79.8	71.0	46	26.1

majority class frequency ranges from 50 to 70%), while the information score ranges between 23 and 50%. The rule sets for different taxa comprised 10 to 20 rules, the average rule length was less than five conditions, and a rule covered on average 15 to 45 examples.

While the performance (accuracy) of the rules on the training set is not as high as might be expected, we should bear in mind that the use of a high significance threshold prevents overfitting. More importantly, the physical and chemical parameters at a certain point in time do not determine completely the presence (absence) of a particular taxon: the presence depends on the physical and chemical parameters over a longer period of time, on the life time of the taxon, the water level, and the river bed. To make the problem even harder, some taxa group together very different organisms: an example is the taxon *Chironomidae (green)*, where the lowest information score on the whole dataset was recorded.

The information scores of the rules induced from 70% of the dataset (measured on the remaining 30%) is much lower for all taxa, but remain positive. This means that the rules contain useful information about the influence of physical and chemical parameters on the presence of the taxa. Nevertheless, the accuracy is worse than the default for 5 out of the 17 taxa. In the remainder of the section, we give an excerpt from the expert evaluation of the rules for the diatom *Nitzschia palea* and the water beetle *Elmis sp.*. The rules for these taxa achieved the highest information scores (29.2% and 26.1%, respectively) on unseen cases (average of the 10 splits 70% training–30% testing).

The diatom *Nitzschia palea* is present in 420 of the 698 samples and is the most common species in Slovenian rivers. It is very tolerant to pollution and lives in waters of a wide quality range, from clean to polluted waters. It is characteristic of the water quality class 2.–3. according to the saprobic index and is used as an indicator of polluted waters.

The rules built from the whole dataset confirm that a larger degree of pollution is beneficial to this species. From the 18 rules, Table 7 lists six rules, chosen to have large coverage. The rules indicate that *Nitzschia palea* needs nitrogen compounds, phosphates, silica, and larger amounts of degradable matter (COD and BOD).

The beetles *COLEOPTERA*, where the taxon *Elmis sp.* belongs, are quite common on land but rare in water. From the literature and expert experiences it is known that this taxon inhabits clean waters: it is considered an indicator of the quality class 1.–2.

From the 17 rules induced, five selected by the expert are listed in Table 8. The first rule demands a relatively low quantity of biodegradable matter (pollution) in order for *Elmis sp.* to be present; this has to be even lower as water temperature increases (see the second and the third rule). The last two rules predict that the taxon will be absent if the water is overly polluted as indicated by the high values of BOD, COD and pH. The rules confirm that *Elmis sp.* is a bioindicator of clean to mildly polluted waters.

Not all of the induced rules agree with existing expert knowledge. As an example, let us consider the rules that predict the presence of the taxon *Plecoptera leuctra sp.*, which is used as an indicator of

Table 7
Selected rules predicting the presence of the species *Nitzschia palea* from physical and chemical parameters. The numbers in square brackets are the numbers of examples of each class covered by the respective rule. For example, [58 0] means that the corresponding rule covers 58 examples of class `Present`, while [0 39] means that the corresponding rule covers 39 examples of class `Absent`

IF	PO4 > 0.065	IF	NO3 > 1.3
AND	Fe < 0.595	AND	NH4 < 0.97
AND	COD > 25.5	AND	13.25 < COD < 16.35
THEN	Taxon = Present [58 0]	THEN	Taxon = Present [36 0]
IF	4.25 < NO3 < 12.35	IF	Hardness > 11.85
AND	SiO2 > 1.65	AND	NO2 > 0.095
AND	Detergents > 0.055	AND	NH4 > 0.09
THEN	Taxon = Present [50 0]	THEN	Taxon = Present [82 0]
IF	NO3 < 5.95	IF	NO2 < 0.005
AND	SiO2 > 4.75	AND	NO3 < 7.1
AND	COD > 7.95	AND	PO4 < 0.125
AND	1.3 < BOD < 42.05	AND	Detergents < 0.055
THEN	Taxon = Present [59 0]	AND	BOD < 2
		THEN	Taxon = Absent [0 39]

Table 8
Selected rules predicting the presence of the taxon *Elmis sp.* from physical and chemical parameters. The numbers in square brackets are the numbers of examples of each class covered by the respective rule. For example, [36 0] means that the corresponding rule covers 36 examples of class Present, while [0 72] means that the corresponding rule covers 72 examples of class Absent

```
IF    O2 < 11.45              IF    Temperature >
                                    12.75
AND   Hardness >              AND   BOD < 0.65
      10.35
AND   COD > 2.15              THEN  Taxon = Present
AND   BOD < 1.25                    [8 0]
THEN  Taxon = Present
      [36 0]                  IF    23 < 46.45
                              THEN  Taxon = Absent
IF    Temperature >                 [0 72]
      11.75
AND   12.3 <
      Hardness < 14.3
AND   BOD < 1.75              IF    PH > 7.05
THEN  Taxon = Present         AND   BOD > 12.15
      [14 0]                  THEN  Taxon = Absent
                                    [0 47]
```

clean waters. The induced rules do confirm that it is found mainly in clean waters. However, they also state that *Plecoptera leuctra sp.* can be found in quite polluted water, provided there is enough oxygen. Thus, they enhance current knowledge on the bioindicator role of this taxon.

Another example are the rules that predict the presence of the taxon *Cymbella sp.*. The rules point out that the taxon can be found in moderately to critically polluted waters (as indicated by the tolerance of large quantities of biodegradable matter, i.e., large values of BOD). In water monitoring practice, however, *Cymbella sp.* is used as an indicator of clean to mildly polluted waters.

4.2. Biological classification

This section describes the experiments in predicting the biological water quality class of Slovenian rivers, as determined by the saprobic index, from two different sets of attributes. The first set consists of all the physical and chemical parameters mentioned in the previous section, altogether 13 parameters. The second consists of the 17 taxa from the previous section and 10 additional taxa, altogether 27 taxa. The 13 parameters give rise to real valued attributes, while the 27 taxa give rise to discrete valued attributes with four (linearly ordered) values: 0, 1, 3, and 5. As mentioned earlier, there are seven water quality classes. The majority class 2. comprises 339 of the 698 examples, thus the default accuracy is 48.6%.

For illustration, let us take a look at the two rules in Table 9 that predict the water quality class from physical and chemical parameters: these are the rules with greatest coverage derived from the whole dataset. While we would need expertise in both the chemistry and biology of water quality to thoroughly evaluate these rules, they are intuitive and understandable. The class 1.–2. (first rule) requires relatively cold water and very small quantities of pollutants (NO_2, NO_3, detergents, COD, BOD). Class 2. waters (second rule) are usually warmer and somewhat larger quantities of pollutants are allowed, provided there is enough oxygen (Saturation > 57.3).

Table 9
Two rules that predict the water quality class from physical and chemical parameters. The first covers 9 examples of class 1., 80 examples of class 1.–2., and 2 examples of class 2. It predicts class 1.–2. The second covers 8 examples of class 1.–2., 152 examples of class 2., and 5 examples of class 2.–3. It predicts class 2.

```
IF        Temperature < 14.35
AND       PH < 8.45
AND       NO2 < 0.235
AND       1.75 < NO3 < 7.15
AND       Detergents < 0.025
AND       COD < 4.25
AND       BOD < 2.35
THEN      QualityClass = 1.-2.
          [9 80 2 0 0 0 0]

IF        Temperature > 12.65
AND       PH < 8.65
AND       Saturation > 57.3
AND       NO2 < 0.375
AND       NH4 > 0.065
AND       PO4 < 0.39
AND       SiO2 < 10.75
AND       COD > 2.65
AND       1.25 < BOD < 4.75
THEN      QualityClass = 2.
          [0 8 152 5 0 0 0]
```

Table 10
Two rules that predict the water quality class from the density levels of selected bioindicators. The first covers 1 example of class 1., 16 examples of class 1.–2., and 1 example of class 2. It predicts class 1.–2. The second covers 3 examples of class 1.–2., 32 examples of class 2., 9 examples of class 2.–3., 2 examples of class 3., and 1 example of class 4. It predicts class 2.

```
IF      BACILLARIOPHYTA_Navicula_
        cryptocephala = 0
AND     CHLOROPHYTA_Scenedesmus_obliquus = 0
AND     DIPTERA_Chironomidae_green = 3
AND     COLEOPTERA_Elmis_sp. = 3
THEN    QualityClass = 1.-2.
        [1 16 1 0 0 0 0]

IF      BACILLARIOPHYTA_Navicula_
        cryptocephala = 1
AND     BACILLARIOPHYTA_Nitzschia_palea = 1
THEN    QualityClass = 2.
        [0 3 32 9 2 0 1]
```

The rules induced on the entire dataset reach 81.5% classification accuracy when using physical and chemical parameters and 71.1% when using bioindicators, the information scores being 62% and 44%. When learning on 70% of the dataset, the corresponding accuracies on the testing set are 60% and 58%, the information scores being 32% and 28%. It is interesting to note that better performance is achieved when predicting from physical and chemical parameters, despite the fact that biological quality is predicted. However, to determine the quality class a much larger set of bioindicators is used than the one used in our experiments.

Let us finally take a look at the two rules in Table 10 that predict the water quality class from the 27 bioindicators. The first rule predicts class 1.–2. when *Elmis sp.* and *Chironomidae (green)* occur frequently (3) and the species *Scenedesmus obliquus* and *Navicula cryptocephala* are absent (0). It is in agreement with existing expert knowledge: *Elmis sp.* and *Chironomidae (green)* are indicators of clean waters, while *Navicula cryptocephala* is indicative of polluted waters (class 3.). The second rule predicts class 2. when *Navicula cryptocephala* and *Nitzschia palea* occur incidentally (1). Both species are indicative of heavily polluted waters if they occur in larger quantities: as they only occur incidentally, the rule is in agreement with expert knowledge.

5. Related work

Few related work has been reported. In general, ready made expert system shells are frequently used. Some of these include rule induction capabilities. Meszaros et al. (1990) used the expert system shell VP-EXPERT to construct a model for predicting algal biomass one week in advance in a lake. They used data originating from a dBASE-IV database to which they applied the rule-induction capabilities of VP-EXPERT. The induced rules were then used in a Pascal program that performed additional calculations and graphical data presentation.

In the domain of water quality prediction, Recknagel et al. (1994) have developed a hybrid expert system to predict water quality in a lake or a reservoir. They used the NEXPERT OBJECT shell as a platform for the implementation in which a hypertext package ToolBook resides and is used (1) as a geographic information system (GIS) interface for data representation, (2) as an intelligent front-end to handle the simulation model and a historical data base, and (3) as an interface to consult knowledge bases on water quality (eutrophication, algal blooms, and pathogens). The knowledge bases were not constructed automatically, but were written in Prolog by experts. The simulation models (a conceptual, a statistical, and fuzzy one) within the shell were constructed manually, too. Although this is an expert system application to ecology, no machine learning is involved.

In the domain of water quality classification, several studies exist for British rivers. These include the studies by Walley et al. (1992) and Ruck et al. (1993). Walley et al. use Bayesian inference, while Ruck et al. use neural networks to perform biological classification of river water quality. An extension of this work is a comparison of several approaches, i.e. Bayesian, neural networks and machine learning methods on the problem of water quality classification (Walley and Džeroski, 1995).

Guerrin (1991) addresses the task of interpretation of measurements taken from an aquacultural ecosystem with artificial intelligence techniques. This bears some similarity to the problem of biological classification of water quality, where biological samples have to be interpreted in terms of their quality class. However, Guerrin formulates the knowledge neces-

sary for interpretation himself, using a particular qualitative formalism developed specifically for that problem. On the other hand, we extract the knowledge necessary for interpretation/classification of biological samples from examples. In contrast to Guerrin, the knowledge is represented in the well-known formalism of IF–THEN rules.

6. Discussion

We have used rule induction to address several ecological problems. These include water quality classification and analysis of water quality data. Data were analyzed and models/rules generated for both British and Slovenian rivers.

For the British rivers, expert classifications of the biological samples were available. The induced rules successfully captured the knowledge of the expert. The expert also found that the rules are generally consistent with his knowledge. Information on the diversity of the population further increased the information content and the quality of the induced rules. We evaluated the performance of the induced rules on unseen cases and demonstrated that they perform better than the kNN classification algorithm, providing further evidence that the rules embody useful generalizations of the training examples. The induced rules/model can be used to automate the task of interpretation of biological samples in water quality terms, i.e., to predict the water quality at a particular site given the structure of the invertebrate community at that site.

Concerning Slovenian rivers, the experiments we have performed indicate that rule induction can be used to analyze water quality data and discover different kinds of knowledge. We induced rules that describe the influence of physical and chemical properties of the water in Slovenian rivers on the presence of selected living organisms that are currently used as bioindicators of river water quality. Expert evaluation of these rules showed that they do indeed capture useful knowledge, as indicated by their positive information scores. In some cases, the rules just confirmed the expert knowledge of the biology of the bioindicator taxon concerned. In others, they revealed new aspects of the biology of the studied taxon, which extend existing knowledge without conflicting it. There were even cases when the rules indicated that the given taxon is used as an indicator for a wrong class of biological water quality.

While the above analysis concerned 17 taxa with relatively well known biology that are routinely used as bioindicators of water quality, we have still been able to find some new knowledge on the biology of the taxa studied and their bioindicator roles. This means that the use of rule induction can improve our knowledge about riverine biology and ecology. A promising direction for further work is thus to extend the analysis to taxa about which relatively little is known and which are currently not used as bioindicators. This could contribute new knowledge both to biology and to the practice of water quality monitoring, as some of the taxa analyzed may turn out to be very good bioindicators.

The induced rules can be used to predict the presence of taxa given a set of physical and chemical parameters. They embody 'shallow' knowledge that can be only used for a particular purpose, in this case for prediction. 'Deep' knowledge (Bratko et al., 1989) can be used for simulation, diagnosis and prediction: it can be also 'compiled' into shallow knowledge by using machine learning techniques. However, 'deep' knowledge is usually explicitly formulated by experts. In our application, we have derived 'shallow' knowledge automatically using machine learning techniques and without 'deep' knowledge to start with. It may even be possible to derive 'deep' knowledge using appropriate machine learning techniques, an appropriate formulation of the learning problem and some prior knowledge about the ecological system studied. This is an important topic for further work.

We also induced rules that predict the river water quality class (as provided through the saprobic index) of Slovenian rivers. The rules that use bioindicator data to this end are mainly consistent with existing expert knowledge: this is understandable, as bioindicator data (albeit on a much larger set of indicators) is used to derive the saprobic index. The rules that predict the biological quality class from the physical and chemical water properties are surprisingly accurate and informative and deserve a more detailed further analysis by experts fluent in both biological and chemical aspects of water quality. It

would be reasonable to induce classification rules that use both bioindicator and chemical/physical data, as the two are complementary to a certain degree.

Regarding the machine learning techniques used, we should note that the rule induction system CN2 is biased towards general discriminative rules. This means that it looks for the shortest (most general) rules that are sufficient to distinguish a particular class from the other classes, rather than the most specific rules that characterize the class. This means, for example, that only an upper or a lower bound on the value of a continuous attribute will be imposed, rather than both, if that is sufficient to discriminate between examples of different classes. Other algorithms that look for most specific rules and have incremental facilities (i.e., can refine a rule set when new examples arrive), such as AQ15 (Michalski et al., 1986) may be used in further experiments, as well as systems for machine discovery (Langley and Żytkow, 1989) and inductive logic programming (Džeroski, 1996, Lavrač and Džeroski, 1994).

Water quality has thousands of dimensions, not all of which are relevant to all users. A general water quality index must therefore, by virtue of its very nature, be subjective. Thus subjectivity cannot be eliminated, only minimized. Its minimization requires that it should only enter into the methodology where it cannot be avoided, i.e. in relation to the target classifications. However, most of the traditional methods for water quality classification introduce subjective judgements at intermediate stages in their development. For example, the allocation of scores in the BMWP system and the allocation of saprobic values and indicator values in the saprobic system (De Pauw and Hawkes, 1993) are subjective acts. In addition, these systems use methods of combining the evidence which are not soundly based theoretically, but merely ad hoc procedures based on summations, averages or weighted averages. Thus both the numbers used and the methods of combining them are subjectively derived.

For the biological classification of British rivers, our approach has been to introduce subjectivity at one point only, the expert classification of a representative cross-section of samples so as to provide a set of examples. We then use rule induction to map directly from the samples to the classifications, with no (or very little) introduction of further subjectivity. The analysis of the influence of physical and chemical parameters on selected organisms in Slovenian rivers helps to understand the biology of these organisms and make the assignment of saprobic values and indicator values more objective. In both cases, rule induction contributes towards better, more objective methods for water quality classification.

In summary, we have described a technique for rule induction and its application to several hydroecological systems. The models produced (in the form of rules) are transparent and can be easily understood by experts. In all the domains the induced rules contained valuable knowledge about the problem studied. In some cases, this knowledge extended and complemented existing expert knowledge. Machine learning techniques are therefore useful tools for ecological modelling, especially in the early exploratory stages.

Acknowledgements

The authors wish to thank H.A. Hawkes for his invaluable assistance in classifying the biological samples and providing expert comments on the rules produced, and also the NRA (Severn–Trent Region) for providing the biological data on British rivers. We would like to thank Doris Ličan-Miloševič, who performed the experiments with the data on Slovenian rivers as a part of her BSc Thesis at the Faculty of electrical engineering and computer science, University of Ljubljana, Slovenia. Her thesis was supervised by Professor Ivan Bratko, Sašo Džeroski and Jasna Grbović. We acknowledge the support of the Hydrometeorological Institute of Slovenia that provided the water quality data on Slovenian rivers, as well as the support of the Slovenian Ministry of Science and Technology.

References

Bratko, I., Mozetič, I. and Lavrač, N., 1989. KARDIO: A Study in Deep and Qualitative Knowledge for Expert Systems. MIT Press, Cambridge, MA.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont.

Cestnik, B., 1990. Estimating probabilities: A crucial task in

machine learning. In: Proc. Ninth Eur. Conf. on Artificial Intelligence. Pitman, London, pp. 147–149.

Clark, P. and Boswell, R., 1991. Rule induction with CN2: Some recent improvements. In: Proc. Fifth Eur. Working Session on Learning. Springer, Berlin, pp. 151–163.

Clark, P. and Niblett, T., 1989. The CN2 induction algorithm. Mach. Learn., 3(4): 261–283.

Dasarathy, B.V., ed., 1990. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Los Alamitos, CA.

De Pauw, N. and Hawkes, H.A., 1993. Biological monitoring of river water quality. In: Proc. Freshwater Europe Symp. on River Water Quality Monitoring and Control. Aston University, Birmingham, pp. 87–111.

Džeroski, S., 1996. Inductive logic programming and knowledge discovery in databases. In: ed. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge, MA, pp. 118–152.

Džeroski, S., Cestnik, B. and Petrovski, I., 1993. Using the *m*-estimate in rule induction. J. Comput. Inf. Technol., 1: 37–46.

Grbović, J., 1994. Applicability of various procedures for the assessment of quality of Torrential streams. PhD Thesis, Biotechnical Faculty, University of Ljubljana, Slovenia, (in Slovenian).

Guerrin, F., 1991. Qualitative reasoning about an ecological process: interpretation in hydroecology. Ecol. Modell., 59: 165–201.

Harmon, P., Maus, R. and Morrissey, W., 1988. Expert systems: Tools & Applications. John Wiley, New York.

Holte, R., Acker, L. and Porter, B., 1989. Concept learning and the problem of small disjuncts. In: Proc. Tenth Int. Joint Conf. on Artificial Intelligence. Morgan Kaufmann, San Mateo, CA.

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.

Kompare, B., Bratko, I., Steinman, F. and Džeroski, S., 1994. Using machine learning techniques in the construction of models. Part I: Introduction. Ecol. Modell., 75–76: 617–628.

Kompare, B. and Džeroski, S., 1994. Two artificial intelligence methods for knowledge synthesis from environmental data. In: ed. P. Zannetti, Computer Techniques in Environmental Studies V (Proc. Fifth International Conference on the Development and Application of Computer Techniques to Environmental Studies), Vol. II: Environmental Systems. Computational Mechanics Publications, Southampton, pp. 265–272.

Kononenko, I. and Bratko, I., 1991. Information-based evaluation criterion for classifier's performance. Mach. Learn., 6(1): 67–80.

Langley, P. and Żytkow, J., 1989. Data-driven approaches to empirical discovery. Artif. Intell. 40: 283–312.

Lavrač, N. and Džeroski, S., 1994. Inductive Logic Programming: Techniques and Applications. Ellis Horwood, Chichester.

Ličan-Miloševič, D., 1994. Analysis of water quality data by rule induction. BSc Thesis, Faculty of electrical engineering and computer science, University of Ljubljana, Slovenia, (in Slovenian).

Meszaros, F., Varis, O., Sirvio, H. and Kettunen, J., 1990. A rule-based water quality model for PC-environment. Environ. Software, 5(3): 158–162.

Michalski, R., Mozetič, I., Hong, J. and Lavrač, N., 1986. The multi-purpose incremental learning system AQ15 and its testing application on three medical domains. In: Proc. Fifth Natl. Conf. on Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pp. 1041–1045.

Official, J., 1978. Regulations on water classification for interstate streams, international waters and coastal waters of Yugoslavia. Off. J. SFRY, No. 6.

Pantle, R. and Buck, H., 1955. Die biologische Überwachung der Gewas und die Darstellung der Ergebnisse. Gas Wasserfach, 96: 603.

Recknagel, F., Petzoldt, T., Jaeke, O. and Krusche, F., 1994. Hybrid expert system DELAQUA – a toolkit for water quality control of lakes and reservoirs. Ecol. Modell., 71: 17–36.

Ruck, B.M., Walley, W.J. and Hawkes, H.A., 1993. Biological classification of river water quality using neural networks. In: Proc. Eight Int. Conf. on Artificial Intelligence in Engineering. Elsevier, pp. 361–372.

Walley, W.J., Boyd, M. and Hawkes, H.A., 1992. An expert system for the biological monitoring of river pollution. In: Proc. Fourth Int. Conf. on the Development and Application of Computer Techniques to Environmental Studies. Elsevier, Amsterdam, pp. 1030–1047.

Walley, W.J. and Džeroski, S., 1995. Biological monitoring: A comparison between Bayesian, neural and machine learning methods of water quality classification. In: Proc. Int. Symp. on Environmental Software Systems. Malvern, PA, forthcoming.

Wettschereck, D., 1994. A study of distance-based machine learning algorithms. PhD Thesis, Department of Computer Science, Oregon State University, Corvallis, OR.

Zelinka, M. and Marvan, P., 1961. Zur Präzisierung der biologischen Klassifikation der Reinheit fliessender Gewässer. Arch. Hydrobiol., 57: 389–407.