Classification of river water quality using machine learning

S. Džeroski, L. Dehaspe, B.M. Ruck & W.J. Walley

a Institute Jožef Stefan, Jamova 39, 61111 Ljubljana, Slovenia
b Department Computerwetenschappen, Katholieke Universiteit Leuven, Belgium
c Department of Civil Engineering, Aston University, Birmingham, UK
d School of Computing, Staffordshire University, Stafford, UK

Abstract

We use machine learning to find rules for the direct interpretation of biological samples of river life. Given are approximately three hundred field samples of benthic communities taken from the upper Trent catchment in England, and classified into water quality terms by an expert river ecologist (H. A. Hawkes). From these, we induce classification rules that capture the expertise of the ecologist. Different machine learning techniques are employed to this end, including rule induction in propositional and first-order logic. The rules were inspected by the ecologist and found to be mostly consistent with his knowledge.

Biological classification of water quality

Biological monitoring of river water quality is of increasing importance [2]. The data considered in this paper consist of 292 samples of benthic macroinvertebrates collected as part of a biological monitoring programme. 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 on a scale from 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. The expert classified the samples 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 [8]. 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 [7]. Class B1a indicates least polluted (best quality) water, while class B4 represents water of the poorest quality.

Sample 74698: 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: Hydrobiidae (4), Physidae (3), Lymnaeidae (1), Ancylidae (2), Sphaeriidae (3), Glossiphoniidae (2), Erpobdellidae (2), Hydracarina (3), Gammaridae (1), Hydropsychidae (4), Tipulidae (2).

Two samples are given above. Sample 74698 was classified by the expert into class B1a and sample 75892 into class B1b. A section of the site by species matrix corresponding to these samples is given below.

<table>
<thead>
<tr>
<th>SampleID</th>
<th>Planariidae</th>
<th>Hydracarina</th>
<th>Gammaridae</th>
<th>Tipulidae</th>
</tr>
</thead>
<tbody>
<tr>
<td>74698</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>75892</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

The process of classification of river water quality could be automated by eliciting the expert knowledge and embodying it in an expert system. The classical way to acquire the knowledge is to interview the expert and formalize his relevant statements in a suitable knowledge representation formalism. Unfortunately, this approach is extremely time consuming and error-prone. Experts find it much easier to perform a task (e.g., classification of biological samples) than to explain how exactly they perform it.

The goal of our work is to capture the expert knowledge directly from the examples classified by the expert. To this end, we apply several machine learning techniques for rule induction in propositional and first-order representations. From the given classified samples, we construct classification rules that capture the expert's underlying knowledge. These rules could be used as a knowledge base for an expert system.
Machine learning

Machine learning is an area of artificial intelligence concerned with using existing experience to improve the process of solving problems such as classification and prediction. A variety of machine learning techniques exist, including induction of classification and regression trees, rule induction, and machine discovery of empirical laws. A recent development in machine learning is the subarea of inductive logic programming [5], which deals with learning in a first-order formalism.

Rule induction techniques, such as CN2 [1], accept as input a set of examples, each described by a fixed set of features, called attributes. The classification of each example into one of a predefined set of classes is also given. From the examples and their classifications, a set of classification rules is induced that capture the regularities in the given examples.

The induction process searches the space of possible rules in a heuristic manner. Usually, not all valid rules (existing regularities) are found, but rather a subset of valid rules which suffice for correct classification of the given examples. These are not necessarily mutually exclusive, i.e. an example from the learning set may correspond to several rules.

In the case of biological classification of water quality, an example is a biological sample. The attributes are the abundance levels of the eighty families of benthic macroinvertebrates, i.e. the columns of the site by species matrix (Planariidae, ..., Hydracarina, Gammaridae, ..., Tipulidae). The class is the water quality class, with possible values: Bla, Bib, B2, B3, and B4.

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

The above rule was induced from the classified biological samples. It states that a sample should be classified into the best quality class B1a if less than 50 animals (<= 3) of the family Hydrobiidae are present, Planorbidae is absent (<= 0), there are less than 1000 animals of the family Gammaridae (<= 5), and there are some animals of the family Leuctridae (> 0). The rule covers 42 examples of class B1a and no examples of the other four classes (numbers between square brackets).

Inductive logic programming (ILP) employs a different, logic-based, representation of the examples and the learned rules. While the examples are represented by a fixed length vector of attribute values in rule induction, they are represented by a number of relational facts in ILP. In addition, ILP allows for the use of existing domain knowledge in the form of background relations.
Consider the sample 75892 mentioned above. In the ILP framework, it can be described by the facts blb(75892), hydrobiidae(75892,4), physidae(75892,3), lymnaeidae(75892,1), ancyliidae(75892,2), sphaeriidae(75892,3), glossiphoniidae(75892,2), erpobdellidae(75892,2), hydracarina(75892,3), gammaridae(75892,1), hydropsychidae(75892,4), and tipulidae(75892,2). The first fact gives the classification of the sample. For each family present in the sample, there is a fact giving its abundance level. The fact tipulidae(75892,2), for example, mean that the family Tipulidae is present at abundance level 2 (3-9 members) in sample 75892.

Rules have the form of first-order clauses in ILP. For example, the rule b1a(X) IF rhyacophilidae(X,A) AND perlodidae(X,B) states that a given sample X belongs to class B1a if both Rhyacophilidae and Perlodidae are present in the sample (at abundance levels A and B). The rule b1a(X) OR b1b(X) IF perlodidae(X,A) states that sample X belongs to either class B1a or class B1b, given that Perlodidae is present in the sample. The variables X, A, B are universally quantified, i.e. the rules hold for all samples X and all abundance levels A, B.

Two main approaches can be distinguished in ILP: the normal semantics approach and the nonmonotonic semantics approach. The normal semantics approach is similar to rule induction, as it is concerned with finding a set of clauses (rules) that classify correctly the given learning set of examples. This goal is achieved through heuristic search of a given space of rules. A representative of the normal semantics approach is the ILP system GOLEM [6], which has been applied to several practical domains.

The nonmonotonic semantics approach to ILP is concerned with finding all rules (clauses), within a given language, that are consistent with the examples. This involves a comprehensive search of the space of rules, which can reveal rules overlooked by the normal semantics approach. In addition, rules with disjunctive conclusions (e.g. b1a(X) OR b1b(X)) are allowed in systems based on the nonmonotonic semantics, but not in those based on the normal semantics (at least at present). A representative of the nonmonotonic approach is the system CLAUDIEN [3].

In the remainder of the paper, we describe the application of the rule induction system CN2, the ILP system GOLEM and the ILP system CLAUDIEN to the problem of learning rules for classification of biological samples into the five river water quality classes.

**Induction of rules for biological water quality classification**

To induce classification rules from the given samples, we applied an extended version of the CN2 system [4], which is able to deal with noisy data
and measures rule performance both in terms of accuracy and information content. The default setting of the learning system were used, except for the search heuristic and the significance threshold. A significance threshold of 99% was employed, enforcing the induction of reliable rules. The heuristic chosen (m-estimate of the rule accuracy with \( m = 32 \)) also directed the search towards more reliable rules.

In addition to the rule given in the previous section, eleven other rules were generated, altogether twelve 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.

The induced rules mention the presence/absence or give other bounds on the abundance levels for 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 [2].

Two more example rules are given below. The first predicts class B2, given the absence of Scirtidae, at least 50 members of Asellidae, and the presence of Gammaridae, which are at the same time required not to be overabundant (less than 100 members). The second rule predicts class B3, relying 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 Glossiphoniididae to be less than 10. 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.

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

IF Planariidae <= 0 AND Tubificidae > 0 AND Lumbricidae <= 0 AND Glossiphoniididae <= 2 AND Asellidae > 0 AND Gammaridae <= 0 AND Veliidae <= 0 AND Hydropsychidae <= 0 AND Simulidae <= 0 AND Muscidae <= 0 THEN Class = B3 [0 0 3 28 10]

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 from the previous section, the expert specified class B1b (actual B1a), for the first rule above B3 (actual B2), and for the second rule above B3 (actual B3).

While the rules were mainly consistent with expert knowledge, some criticism was expressed regarding the reliance of the rules on the absence of families from samples, since their absence may not be significant. In many cases, but not all, the absence of a taxon 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 as to 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 sample. The same settings as above were used, except for the parameter m in the search heuristic (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 is given below. It predicts class B4 (poorest quality). At most five different taxa are allowed in the sample, and at least two of these have to be present with at least three animals.

\[
\text{IF Oligochaeta} \leq 5 \text{ AND Asellidae} \leq 3 \text{ AND Dixidae} \leq 0 \text{ AND MoreThan0} \leq 5 \text{ AND MoreThan1} > 1 \\
\text{THEN Class} = \text{B4 [0 0 0 4 25]}
\]

Applying inductive logic programming to biological water quality classification

The ILP system GOLEM [6] was applied to the problem of classification of biological samples. To suit the normal ILP semantics, the problem was formulated as follows: for each class a separate learning problem was created, where the positive examples are the samples classified in that class, and all the other samples are negative examples. Thus, rules for each of the relations b1a(X), b1b(X), b2(X), b3(X), and b4(X) were generated. The samples were described by eighty predicates of the form family(X,A), denoting that family is present in sample X at abundance level A. Relations of this kind include tipulidae(X,A), asellidae(X,A), etc. In addition, the
background relation `greater_than(A,B)` was available, stating that abundance level A is greater than abundance level B.

The default settings of GOLEM were used, except for the fact that rules were allowed to cover at most five negative examples. GOLEM produced three rules for class B1a, fourteen rules for B1b, sixteen for B2, two for B3, and none for B4. An example rule is `bla(X) IF leuctridae(X,A)`, which states that a sample belongs to the best water quality class if Leuctridae are present. This rule covers forty-three positive and four negative examples, and agrees with expert knowledge; the family Leuctridae is indicator of good water quality. Another interesting rule is `bla(X) IF baetidae(X,A) AND elminthidae(X,B) AND heptageniidae(X,C) AND perlodidae(X,D) AND limnephilidae(X,E)`, which covers twenty-seven positive examples and is considered a good rule by the expert.

The rule `blb(X) IF ancylidae(X,A) AND gammaridae(X,B) AND hydropsychidae(X,C) AND rhyacophilidae(X,D) AND greater_than(B,A) AND greater_than(B,D)` is considered good. Gammaridae in abundance is a good indicator of class Bib, along with the other families present. The rule `b2(X) IF dytiscidae(X,2) AND gammaridae(X,l) AND lymnaeidae(X,A) AND greater_tha(n(A,1)` is judged to be a fair rule, but a bit too specific. It requires Dytiscidae to be present at abundance level 2 (3-9 members), Gammaridae at level 1 (1-2 members), and Lymnaeidae at a level greater than one, i.e. more than 2 members.

Out of the thirty-five rules, twenty-five are considered good or acceptable by the expert, but some of them are judged to be too specific. Note that GOLEM cannot use the absence of taxa in the rules in the representation adopted. Together with GOLEM's strategy of looking for the most specific rules consistent with the examples, this may have influenced the generality (specificity) of the generated rules. This may also be the reason that no rules for class B4 were induced. Namely, the absence of most taxa is characteristic for this class.

Unlike GOLEM, CLAUDIEN [3] checks all possible rules, within a specified language, for consistency with the given example. Only tests of the presence of different taxa were allowed in the rules (no absence tests and no abundance level tests). Rules were required to cover at least thirty examples.

Altogether, seventy-nine rules were generated. Of these, twenty-eight involved the presence of a single taxon. These rules in fact specify the range of quality classes in which a certain family is present. The rule `bla(X) OR blb(X) IF perlodidae(X,A)` specifies that Perlodidae is found in good quality water (classes B1a and B1b). If Rhyacophilidae is also present, then the water is almost certainly of class B1a: `bla(X) IF perlodidae(X,A) AND rhyacophilidae(X,B)`. Both rules are judged by the expert to be good.

Two thirds of the rules were considered to be good. Only nine rules had a single class (B1a) in the conclusion. The others had a number of possible
classes in the conclusion, which was considered natural and understandable by the expert. This indicates that class B1a is easy to characterize in terms of the families present, while for the other classes references to the abundance levels or absence of certain families is required in order to find strong rules (that cover at least thirty samples).

The rule b1a(X) IF tipulidae(X,A) AND perlodidae(X,B) AND limnephilidae(X,C) AND heptageniidae(X,D) AND baetidae(X,E) is very good. It covers thirty-four samples and mentions the presence of five families. An example of a rule considered somewhat problematic is the following: b1b(X) OR b2(X) IF hydrobiidae(X,A) AND asellidae(X,B). The expert said the conclusion should be B2 or B3, but the rule indicates B1b or B2 with the support of eighty samples. The main reason for the difference seems to be that in the absence of further evidence about the rest of the sample the expert placed greater emphasis on the presence of Asellidae than Hydrobiidae, because Asellidae is recognized as a key indicator.

Discussion

We applied machine learning to generate rules for classification of river water quality based on biological samples. As compared to other methods for automating the classification process, such as neural networks, our approach produces symbolic rules that can be used as a knowledge base for an expert system. The expert was reluctant to provide classification rules explicitly; his expertise was provided in the form of sample classifications and was elicited using machine learning.

Three different machine learning methods were applied. All produced rules that were, to a great extent, consistent with expert knowledge. The rules generated by CLAUDIEN were judged to be the most intuitive and promising. This is due to the coverage of more than one class by a single rule; this is important as the classification problem is based on a discretization of a continuous space. The fact that a single taxon is used in some rules also contributes to the overall intuitiveness of the CLAUDIEN rules.

Rules generated by GOLEM were considered too specific. The CN2 rules were quite general, but relied too heavily on the absence of certain taxa. However, the absence of some taxa seems to be necessary for rules classifying into poorer water quality classes. The best direction for further work seems the selective use of taxa absence. A distinction should be made between the absence of common taxa and rare taxa (which only occur occasionally even in their preferred habitats). The absence of the latter is of no significance, but their presence may be highly significant. So, it is only worth mentioning absentees when they refer to taxa which are normally quite common.

Knowledge of the kind mentioned above, concerning common or rare taxa, can be easily incorporated within CLAUDIEN. This, together with the use of specific abundance levels, comparisons between them and the
use of diversity information, appears to be the most promising direction for further work. The performance of the rules obtained in this way will be tested on an independent data set and compared to the equivalent performance of neural networks and Bayesian systems developed at Staffordshire University. If their performance so merits the rules will then be used as a knowledge base of an expert system for biological classification of river water quality.

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 field data. The research reported in this paper was in part supported by the ESPRIT Project 6020 Inductive Logic Programming.

References


