20

Biological Monitoring: a Comparison between Bayesian, Neural and Machine Learning Methods of Water Quality Classification.

W. J. Walley
School of Computing, Staffordshire University
Beaconside, Stafford ST18 0DG, UK.
Tel: +1785 353510, Fax: +1785 353497
E-mail: W.J.Walley@soc.staffs.ac.uk

S. Džeroski
Institut Jožef Stefan
Jamova 39, 61111 Ljubljana, Slovenia.
Tel: +386 61177 3217, Fax: +386 61125 1038
E-mail: saso.dzeroski@ijs.si

Abstract

Biological methods of monitoring river water quality have enormous potential but this is not presently being realised owing to inadequacies in methods of data interpretation and classification. This paper describes the development and testing of several classification models based on Bayesian, neural and machine learning techniques, and compares their performance with two traditional models. It is demonstrated, using an expertly classified test data set, that 'naive' Bayesian models and multi-layered perceptrons can significantly out-perform the traditional methods. It is concluded that these two techniques presently provide the most promising means of realising the full potential of bio-monitoring, either acting separately or jointly as complementary 'experts'.

Keywords

Bio-monitoring, water quality, classification, Bayesian, neural networks, machine learning

R. Denzer et al. (eds.), *Environmental Software Systems*

1 INTRODUCTION

Biological Monitoring

Biological methods of monitoring environmental quality are of increasing importance. The basic principle of these methods is that the type and diversity of the biota at a given site reflect its environmental quality. The biota act as resident observers of prevailing conditions and provide evidence which can be interpreted into quality terms. Kolkwitz and Marsson (1902) were the first to propose the use of biota as a means of monitoring the quality of natural waters. Since then many different methods of mapping biological data to discrete quality classes or continuous scales have been developed (DePauw and Hawkes, 1993). For practical reasons, the majority of these methods are based exclusively on the composition of the macro-invertebrate community living on or in the bed of the river or lake. Unfortunately, the rationales behind the methods used to interpret the community data are typically *ad hoc* and highly subjective, and their reliabilities have been less than desired. Walley *et al.* (1992) and Ruck *et al.* (1993) first demonstrated the potential of Bayesian and neural classifiers to provide more reliable interpretations, and this present study extends the search for better methods via an appraisal of an even wider range of AI (Artificial Intelligence) techniques.

Basis of the Study

A comparative study of a wide range of computer-based methods of classification (Michie *et al.*, 1994) found that relative performances differed significantly between domains, depending upon the nature of the data. However, none of the data sets used in the study was in any way related to biological monitoring of environmental quality, so no conclusions can be drawn from it as to the best methods to apply in this field.

The aim of this study was to develop and test several computer-based methods of classifying river water quality with a view to identifying the method having greatest potential for future use. This was achieved by comparing their performances relative to traditional methods using test data which had been classified by an expert. The computer-based methods used three different AI techniques: 'naive' Bayesian inference (a probabilistic approach); supervised and unsupervised neural networks and machine learning of regression/model trees (Quinlan, 1993). The traditional methods used were the Average Score Per Taxon (ASPT) and Trent Biotic Index (TBI) as applied in the United Kingdom.

2 THE DATA SET

The raw data set consisted of 292 biological samples taken from sites on the Upper Trent in England. The samples had been classified in water quality terms by an acknowledged expert river ecologist, H. A. Hawkes. This gave the best possible mapping from data to quality class, given the present state of the art. The ASPT's and TBI's were derived by the National Rivers Authority and supplied with the data. Since some of the taxa (i.e. different organisms) found in the samples occurred infrequently or were of little indicator value, a short-list of 36 taxa was drawn up by the expert to form the basis of the project database. The list comprised 7 species, 2 genera, 26 families and 1 order, and was designed to provide effective coverage of the whole range of water quality. Within the samples each taxon was identified as being either: absent, present (1-2 individuals), few (2-9), common (10-49), abundant (50 - 99), very abundant (100 - 1000) or greater than 1000. For the purpose of the neural and machine learning models, these six 'states of existence' were coded in the database as -0.2, 0.2, 0.4, 0.6, 0.7, 0.8 and 0.9 respectively. Thus, the absence of a taxon, being represented by a non-zero value, was able to contribute to the classification. Water quality was represented by five exhaustive and mutually exclusive classes (B1a, B1b, B2, B3 and B4) from the best to the worst quality. These were represented by Classification Indices (CI) from 1 (B4) to 5 (B1a) in the database. However, when classifying the samples the expert was able to identify good (+), normal and bad (-) samples within each class. Thus each class was effectively sub-divided into thirds, so these were represented within the database by CI's differing by 0.33 of a class interval (e.g. B2+ = 3.33). To facilitate independent testing of the models, the database was divided into two representative sub-sets: a *training set* of 195 samples and a *test set* of 97 samples.

3 THE MODELS TESTED

Traditional Methods

The two traditional methods tested in this study, the TBI and ASPT, are typical of the many different systems which are presently in use. They are subjective *ad hoc* systems based on simple arithmetic and/or look-up tables. De Pauw and Hawkes (1993) comprehensively reviewed most of the bio-indicator systems used in Europe, including the TBI, ASPT and Saprobic systems. The TBI was first used by the Trent River Board in England in the late 1950's, and has since formed the basis of indices adapted for use in France, Belgium and Italy. The ASPT is based upon the BMWP score system which was developed in the late 1970's by the Biological Monitoring Working Party, and is unique to the UK. Unfortunately, the Saprobic system, which is used in Germany and many east European countries, could not be tested in this study because its data requirements were not compatible with the project data set.

The TBI is based on a two-way look-up table in which the rows represent different taxonomic groups arranged in decreasing order of pollution sensitivity, and the columns represent increasing levels of the total number of groups present. The user looks up an index (between 0 and 10) at the intersection of the row corresponding to the most pollution sensitive group in the sample and the column corresponding to the total number of groups in the sample. Essentially, it is a subjective pattern recognition system based upon these two features.

In the BMWP system each macro-invertebrate family has been allocated a score (between 1 and 10) indicative of its sensitivity to pollution. Thus pollution sensitive animals like stoneflies score 10, whilst pollution tolerant ones like sludge worms score 1. The BMWP score for a given site is derived by summing the scores of all the families present at the site. However, since this is sensitive to sampling effort and the seasonal behaviour of some of the families, practitioners now favour the ASPT (i.e. BMWP score divided by the number of contributing families). This, being an average, is less sensitive to sampling effort and seasonal effects and is considered the more reliable indicator.

Thus, in theory the ASPT has a continuous scale from 0 to 10, but in practice its range is from about 1.5 to 7.5, corresponding approximately to classes B4 to B1a. The equivalent range for the TBI is 1 to 10. Linear regression on the *test set* was used to map these to the 1 to 5 scale adopted in the study. The resultant mappings were: ASPT/1.28 and TBI/2.00.

Naive Bayes

When Bayesian inference is applied to classification problems involving several mutually exclusive and exhaustive classes it is normal to assume that all items of evidence are conditionally independent. Such models are commonly referred to as 'naive' Bayes, because the assumption is considered naive. However, in bio-monitoring the assumption that the occurrence of each taxon is independent of that of the others, given the water quality class, appears to hold for the vast majority of taxa. Thus the 'naive' Bayes model appears well-suited to this domain.

Two 'naive' Bayes models have been tested, differing only in the way in which their prior and conditional probabilities are derived. The first used a well-founded statistical method, the *m-estimate* (Cestnik, 1990), to estimate the 'true' conditional probabilities from their raw values. In his case:

$$P'(H_i|e_{jk}) = \frac{n_{ijk} + mP(H_i)}{N_{jk} + m} \tag{1}$$

where $P'(H_i|e_{jk})$ = the revised conditional probability that the water quality H is class i, given evidence e_{jk} that taxon j is in state k.
n_{ijk} = number of occurrences in the data of taxon j in state k in class i ;
N_{jk} = total number of occurrences of taxon j in state k across all classes;
$P(H_i)$ = the prior probability of class i derived from the data set; and
m = a parameter of the estimation method.

The second was strictly based on the Principle of Indifference and used a simple *ad hoc* procedure to eliminate zeros from the raw conditional probabilities, by uniformly redistributing a small percentage of the raw distributions. In this case:

$$P'(H_i|e_{jk}) = (1-x)P(H_i|e_{jk}) + \frac{x}{M} \tag{2}$$

where: x = fraction of total probability redistributed;
M = number of classes;
$P(H_i|e_{jk})$ = $P(e_{jk}|H_i)/\Psi$, where Ψ = the normalisation constant for i = *1 to M;*
(NB. This follows from the application of the Principle of Indifference and the exhaustive nature of the classes i = *1 to M)*
$P(e_{jk}|H_i)$ = n_{ijk}/N_i, where N_i = number of occurrences of class i in the data.

The significance of the parameters m and x is briefly discussed later Five different models of each type were tested: m = 0, 1, 2, 5 and 10; and x = 0, 0.001, 0.05, 0.10 and 0.15.

Neural Nets

Two types of neural network were tested: multi-layered perceptrons (MLP's) which use supervised learning based on the back propagation of errors; and self organised maps (SOM's) which use unsupervised learning based on Kohonen's feature mapping algorithm. The interested reader is referred to the text by Beale and Jackson (1990), which provides a good introduction to neural computing and includes details of the structure and training algorithms of the two networks used here.

The MLP's had 36 input nodes (one for each taxon), one output node and either one, two or three hidden layers. Three networks had just one hidden layer (MLP60, MLP80 and MLP120) with six, eight, and twelve nodes respectively. Two had two hidden layers: MLP63 with six nodes in the first layer and three in the second, and MLP84 with eight in the first and four in the second. The hyperbolic tangent transfer function was used in all cases and training was achieved using the Quick Propagation algorithm. The inputs were not re-scaled, thus their range remained -0.2 to +0.9. The target values for the single output node in the training mode were the CI values of the samples, which ranged from 1 to 5, whilst the predicted CI's during testing were found to range from 0.86 to 5.17.

The SOM's were single layer networks in which the 36 input nodes mapped onto a rectangular feature map. These networks learn by recognising different patterns in the data and allocating each, in terms of its feature vector, to a specific output node in the feature map. This is achieved in a way such that adjacent nodes represent similar patterns, thus producing a matrix of gradually varying feature vectors, hence the term feature map. Since training is achieved without any knowledge of the samples' classifications, the pattern represented by each node has to be identified and labelled to give it meaning. This was easily achieved in this study by examining the expert's prior classifications of the samples allocated to each node. Once a network is trained and labelled, it classifies samples by allocating them to a node in the feature map on the basis of the best match between their features and the node's feature vector. For the purpose of this study, three nets were trained: SOM1x10, SOM3x8 and SOM5x5 having 1x10, 3x8 and 5x5 feature maps respectively.

Regression and Model Trees

Regression trees (Breiman *et al.*, 1984) represent a functional dependency between a continuous valued class and several independent variables, which is not uniform across the whole domain. The leaves of the tree correspond to sub-domains where the dependency can be approximated to a constant (regression tree) or linear model (model tree). The nodes of the tree correspond to independent variables and the branches to tests on these variables. The M5 system for the induction of model trees (Quinlan, 1993) was used to build the model tree (MT1) shown in Figure 1 using the 36 attributes of the data set (i.e. the indicator taxa). To understand how the tree should be interpreted, consider a sample in which Heptageniidae is present, Leuctridae is absent and Nemouridae is few (i.e. the corresponding attribute has a value of 0.2). Moving through the tree, we end up at the leaf in which the Classification Index, CI = 4.26 + 1.53 Nemouridae, and obtain the prediction that CI = 4.26 + 1.53 x 0.2 = 4.87.

Since MT1 was found to be not very accurate, giving only 58.8% correct classifications, M5 was provided with six additional attributes that describe the diversity of the community. The result was the model tree (MT2) depicted in Figure 2. This tree is smaller but more accurate than MT1. Only one of the six attributes, the number of taxa present, was found to be relevant. In fact, it turned out to be the most relevant attribute.

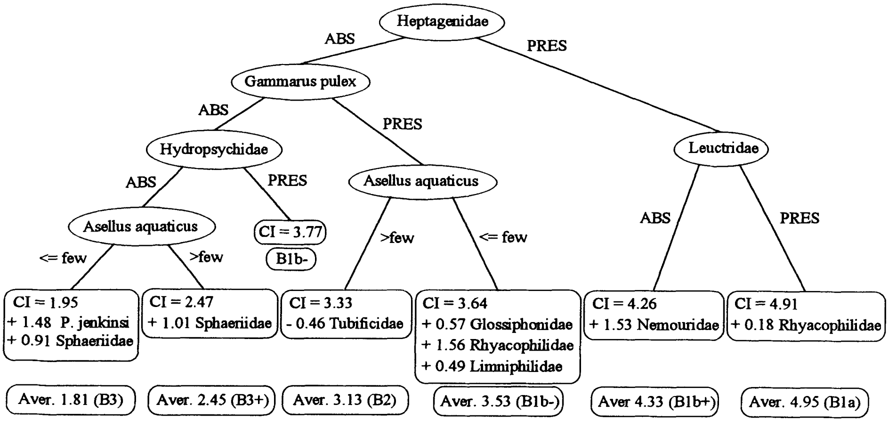

Figure 1 Model Tree No.1 (MT1), derived from 36 attributes.

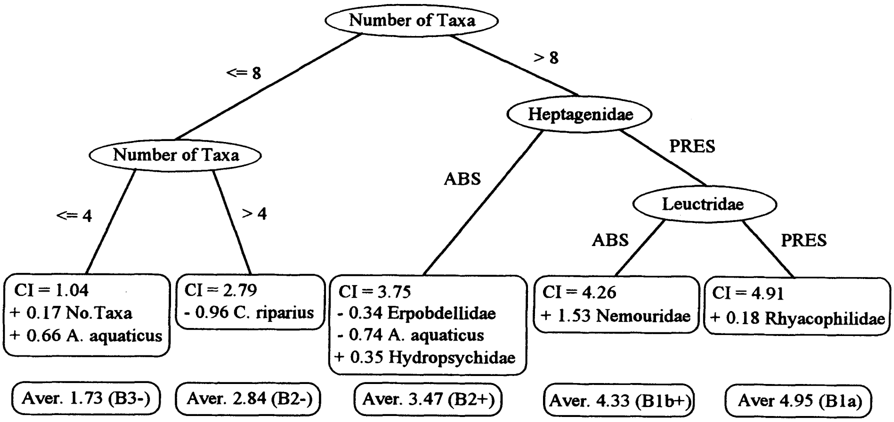

Figure 2. Model Tree No.2 (MT2), derived from 42 attributes.

4 RESULTS OF PERFORMANCE TESTS

Two different performance measures were used to assess the performance of the various models: their classification rates and uniformity of precision across the classes. These were further divided into two types of precision: a) the percentage of correct categorical classifications into classes B1a, B1b, B2, B3 and B4; b) the percentage of predicted CI values lying within 0.25 (narrow-band criterion) and 0.50 (broad-band criterion) of a class interval. Since the outputs of most of the models were in terms of the CI scale, the boundaries between the discrete classes were taken to lie at the 0.5 points on the scale. This assumption and that underlying the second set of tests strictly requires that the CI scale be an interval scale. That is, any given fraction of a class interval should mean the same in terms of change in water quality anywhere on the scale. However, since the scale was only used to compare like with like and not as an absolute interval scale, this issue clearly had little bearing on the value of the results as a means of comparing the relative performances of the models. Thus it was disregarded.

All models except the Bayesian models produced outputs on the CI scale. The Bayesian outputs were in terms of five probabilities, one for each of the discrete classes. These were converted to a single value on the CI scale by deriving the mean value, from:

$$\mathrm{CI} = \sum_{i=1}^{i=5} iP(H_i) \tag{3}$$

where i is the class number (i.e. 1 for B4 through to 5 for B1a) and $P(H_i)$ is the predicted probability of class i. The overall classification rates of the best performing models in each category are given in Table 1.

Table 1. Overall classification rates for selected models

	Percent.	*Predicted CI's within stated*	
	Correct	*fraction of a class interval (%)*	
	Categ.	*<0.25*	*<0.50*
ASPT/1.28	76.3	49.5	79.4
Bayes (x =0.10)	84.5	71.1	88.7
Bayes (m=1)	83.5	67.0	85.6
Bayes (m=5)	85.6	62.9	86.6
MT2	73.2	53.6	79.4
MLP63	82.5	61.9	88.7
MLP84	81.4	66.0	90.7
SOM3x8	74.2	50.5	77.3
CM84(x=0.1)	82.5	69.1	92.8
CM84(m=5)	83.5	66.0	93.8

Once these figures were known it was decided to combine the best two models on a 50/50 basis to form a 'consensus' model. In fact, two consensus models were formed: CM(84/x=0.1) which combined MLP84 with Bayes(x=0.1); and CM(84/m=5) which combined MLP84 with Bayes(m=5). The classification rates achieved by these are shown at the bottom of Table 1.

Table 2 shows the distribution of correct categorical classifications across the five classes for each of the models listed in Table 1. The standard deviations of the success rates are provided as a measure of the models' variability in performance across the classes. Table 3 shows the equivalent statistics for the broad-band and narrow-band precision tests. The figures under each class heading represent the percentage of samples in the class (i.e. incl. +ve and -ve subclasses) with predicted CI values lying within the stated range of the expert's classification.

Table 2. Distribution of percentage correct categorical classifications across the five quality classes for selected models

						Stnd
	B1a	*B1b*	*B2*	*B3*	*B4*	*Dev.*
ASPT/1.28	68.4	81.8	78.1	66.7	88.9	8.3
Bayes(0.1)	89.5	68.2	93.8	86.7	77.8	9.1
Bayes(m=1)	89.5	63.6	96.9	86.7	66.7	13.1
Bayes(m=5)	84.2	68.2	96.9	86.7	88.9	9.4
MT2	57.9	77.3	81.3	80.0	55.6	11.3
MLP63	100.0	68.2	90.6	80.0	55.6	15.8
MLP84	100.0	63.6	90.6	80.0	55.6	16.5
SOM3x8	84.2	50.0	90.6	53.3	88.9	17.9
CM(84/x=0.1)	89.5	59.1	96.9	86.7	66.7	14.4
CM(84/m=5)	94.7	59.1	96.9	86.7	66.7	15.2
Average	85.3	66.4	91.3	78.0	71.2	

Table 3. Distribution of percentage classifications correct to within <0.25 and <0.50 of a class interval for selected models

	<0.25 of a class interval					*Stnd*	*<0.50 of a class interval*					*Stnd*
	B1a	*B1b*	*B2*	*B3*	*B4*	*Dev.*	*B1a*	*B1b*	*B2*	*B3*	*B4*	*Dev.*
ASPT/1.28	47.4	54.6	46.9	33.3	77.8	14.6	73.7	90.9	81.3	60.0	88.9	11.3
Bayes(x=0.1)	79.0	63.6	84.4	46.7	66.7	13.1	89.5	86.4	96.9	80.0	77.8	6.8
Bayes(m=1)	68.4	45.5	87.5	53.3	66.7	14.4	89.5	77.3	96.9	73.3	77.8	8.8
Bayes(m=5)	68.4	40.9	81.3	46.7	66.7	14.9	84.2	77.3	100.0	73.3	88.9	9.3
MT2	57.9	63.6	56.3	53.3	11.1	19.0	63.2	81.8	84.4	86.7	77.8	8.3
MLP63	79.0	54.6	68.8	40.0	55.6	13.3	100.0	77.3	93.8	86.7	77.8	8.9
MLP84	84.2	59.1	75.0	33.3	66.7	17.3	100.0	77.3	93.8	93.3	88.9	7.6
SOM3x8	78.9	31.8	53.1	40.0	44.0	16.2	84.2	68.2	87.5	53.3	88.9	13.7
CM(84/x=0.1)	84.2	59.1	78.1	46.7	66.7	13.4	89.5	86.4	96.9	93.3	100.0	4.9
CM(84/m=5)	79.0	54.6	78.1	40.0	66.7	14.8	94.7	81.8	100.0	93.3	100.0	6.7
Average	72.6	52.7	70.9	43.5	58.8		86.9	80.5	93.1	79.3	86.7	

5 DISCUSSION

River ecologists are unquestionably the real experts in the field, well able to perform complex mental mappings from data to water quality, but the data interpretation models which they have devised to date are too simplistic to produce adequate levels of performance. Two of the AI techniques tested in this study, 'naive' Bayes and multi-layered perceptrons, significantly out-performed the two traditional methods tested. However, before appraising their performances in detail it is worth considering two domain specific factors.

Domain Specific Considerations

Firstly, categorical classes and continuous indices are both in common use. River quality is commonly defined in terms of discrete classes, normally five as in this study, and these are displayed in different colours on river maps. This is done primarily to facilitate the dissemination of information to the public and politicians in an easily understood form. In an operational setting, however, it is often necessary to be more specific, so that one can determine how quality is changing in time or space, or whether a site is near the top or bottom of its stated class. For these purposes more precise continuous indices of quality are required.

Secondly, the frequency of the quality classes is not uniformly distributed. There are fewer B3 and B4 rivers than there are B1a and B1b, but precision in classification is just as important in poor quality rivers as it is in good ones, perhaps even more so. Consequently, a good model must not only achieve a high overall proportion of correct classifications, but this precision must be fairly uniformly distributed across the classes. Thus the variability of the success rates across the classes, as reflected by their standard deviation, should be as small as possible.

Overall Performance Rankings

In order to achieve an overall measure of relative performance, each model was ranked with respect to its success rate and uniformity of distribution using the results given in Tables 1, 2 and 3. The score, on which the overall ranking was based, was calculated as the weighted sum of the rankings on each test using a weight of two for the classification tests and one for the uniformity tests. It is clear from the results of this analysis (Table 4) that Bayesian inference was the most successful of the AI techniques tested, followed by multi-layered perceptrons.

Table 4. Rankings of models based on classification rates and uniformity of precision

	Ranking on Classif'n Rates				*Rank'g on Unif. of Precision*				*Overall*	
	Categ.	*<0.25*	*<0.5*	*O'all*	*Categ.*	*<0.25*	*<0.5*	*O'all*	*Score*	*Rank'g*
Bayes(x =0.1)	2	1	4	1	2	1	3	1	20	1
CM(84/x=0.1)	6	2	2	3	6	3	1	2	30	2
CM(84/m=5)	3	5	1	2	7	6	2	5	33	3
Bayes(m=1)	4	3	7	6	5	4	6	4	43	4
Bayes(m=5)	1	6	6	4	3	7	8	7	44	5
MLP84	7	4	3	5	9	9	4	9	50	6
MLP63	5	7	5	7	8	2	7	6	51	7
ASPT/1.28	8	10	8	8	1	5	9	3	67	8
MT2	10	8	9	9	4	10	5	8	73	9
SOM3x8	9	9	10	10	10	8	10	10	84	10

The two consensus models CM(84/x=0.1) and CM(84/m=5) were beaten only by the Bayes(x=0.1) model, which was the overall best performer on both the classification and the uniformity rankings. However, the CM's did produce the highest success rates on the broad-band (<0.5) criterion, but there was little evidence that a consensus approach is capable of producing a significant improvement relative to stand alone models. The poor performances of the SOM's and the TBI, which was eliminated at an early stage, were partly due to them being categorical classifiers not well suited to use on continuous scales, but even so they did not perform particularly well as categorical classifiers.

The difficulties which most models had in achieving high performance levels on classes B1a and B3 were largely due to the fact that these two classes had high proportions of samples with + and - classifications. Thus, they had a high percentage of borderline cases and presented a greater challenge to the models.

Traditional Methods

In order to offset the possibility of ASPT and TBI being disadvantaged due to them not having been 'trained' on the training set, they were given the benefit of being calibrated to the CI scale by linear regressions on the test set (section 3). In addition, the expert who classified the samples in the database was a leading figure in the development of the system on which the ASPT is based. In fact, he chaired the working sub-group of the committee which produced the system, so his thinking is 'in tune' with ASPT. Thus the authors believe that, if anything, ASPT was advantaged and not disadvantaged in these tests.

Bayesian Models

Much of the success of the Bayesian models was due to their assumption that the data was unreliable, especially with respect to the total absence of some taxa from some classes. Both of the methods tested, m-estimate and *ad hoc* zero-elimination, effectively eliminated zero probability values from all classes. The impact which this had on performance can be seen from the results of the x=0 (i.e. zeros remain) and x=0.1 (i.e. zeros replaced by 0.02) models, which gave 70.1% and 84.5% correct categorical classification respectively. The effect of the different assumptions underlying these two methods is difficult to discern in the results. Both models performed very well, although Bayes(x=0.1) out-performed Bayes(m=1) and Bayes(m=5) on both classification rates and uniformity of precision. Whether or not this was statistically significant is difficult to say. Clearly, the different assumptions about prior probabilities, $P(H_i)$, must have produced different distributions of precision, thus the application of the Principle of Indifference in Bayes(x=0.1) may have been responsible for its superior uniformity figures. However, this requires further investigation before firm conclusions can be drawn.

Multi-layered Perceptrons

The multi-layered perceptrons did not achieve quite the same levels of performance as the Bayesian models, but they have scope for improvement, both in terms of their internal formulation and the coding their data inputs. Furthermore, they will readily facilitate more complex formulations of the water quality mapping, involving environmental factors, simply by the addition of the appropriate variables to the input vector. Thus, they are considered strong contenders for use in future systems, either acting alone or in conjunction with Bayesian models. The Bayesian and neural network approaches are seen as complementary and thus well-suited for use in systems which take a consensus view.

Model Trees

Although the model tree (MT2) finished below ASPT overall, this was only due to its exceptionally poor performance (11.1%) on just one class (B4) in the narrow-band tests. This resulted in it finishing last in the narrow-band uniformity test, despite achieving excellent uniformity over classes B1a to B3. Although their performance was relatively poor, it is worth noting that the model trees used very few of the 36 indicator taxa (see Figures 1 and 2), a fact which could be of practical value. MT1 used twelve of the taxa, five in the decision tree and seven in the regression equations. MT2, which performed much better than MT1 (e.g. 73.2% to 58.8% on categorical classifications), used only eight of the taxa, two in the decision tree and six in the regression equations. The key to MT2's better performance was its use of the 'Number of Taxa' in the decision tree and the regression equations. In effect, this pooled some of the information from the missing twenty-eight taxa into one statistic, but too much was lost to enable the model to match the Bayesian and MLP models. They had the benefit of using the information provided by all 36 taxa. Nevertheless, the model trees do have one advantage over the other models in that they are transparent. Their reasoning is clearly apparent to the user, and in the case of MT1 and MT2 this reasoning has been confirmed by H. A. Hawkes as being broadly consistent with expert knowledge, albeit somewhat limited in its scope. Given a larger training set, these models may be able to achieve greater sophistication and hence better performance, but at present the benefit of their transparency is insufficient to compensate for their poor performance.

6 CONCLUSION

A variety of computer-based water quality classification models have been developed and tested alongside two traditional models, namely ASPT and TBI. Three different AI techniques were used to develop the computer-based models: 'naive' Bayesian inference, neural networks and machine learning (model trees), and two different formulations of the Bayesian and neural models were used.

The overall best performer was judged to be the Bayes(x=0.1) model, but this was closely followed by two systems CM84(x=0.1) and CM84(m=5) which took the consensus view of a Bayesian model and a multi-layered perceptron. Taken overall, two types of system clearly out-performed ASPT (the best of the traditional models): the Bayesian models and the multi-layered perceptrons, but the former marginally out-performed the latter. However, in view of the scope for further improvement to the multi-layered perceptrons it seems that they, together with the Bayesian models, are the most promising techniques for use in future bio-monitoring systems, either acting alone or together as complementary 'experts'. A cross-validation evaluation of the methods is necessary in order to provide a more reliable estimate of their relative performances.

The key to the realisation of the full potential of bio-monitoring lies in the development of reliable data interpretation and classification systems. Biological data are potentially very rich in information, but the mappings necessary to realise this are complex. This paper has demonstrated the potential of some AI techniques to provide such mappings and has provided pointers to future developments.

7 ACKNOWLEDGEMENTS

The authors wish to thank H. A. Hawkes for his invaluable assistance in classifying the biological samples, and also to the NRA (Severn-Trent Region) for providing the data. Thanks are also due to Ross Quinlan for providing a copy of his M5 system, and to Brendan Ruck for constructing the database.

8 REFERENCES

Beale R. and Jackson T. (1990) *Neural Computation: An Introduction.* Adam Hilger, Bristol and New York.

Breiman L., Freidman 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. European Conference on Artificial Intelligence.* Stockholm, Sweden.

De Pauw N. and Hawkes H.A. (1993) Biological monitoring of river water quality, in *River Water quality Monitoring and Control,* ed. W. J. Walley and S. Judd. Aston University, Birmingham. 87-111.

Kolkwitz R. and Marsson M. (1902) Grundsätze für die biologische Beurteilung des Wassers nach seine Flora und Fauna. *Mitt. Prüfungsanst. Wasserversung. Abwasserbeseit.* **1**. 33-72.

Michie D., Spiegelhalter D.J. and Taylor C.C. (Eds) (1994) *Machine Learning, Neural and Statistical Classification.* Ellis Horwood.

Quinlan J. (1993) Combining instance-based and model-based learning, in *Tenth International Conference on Machine Learning.* Morgan Kaufman. 236-243.

Ruck B.M., Walley W.J. and Hawkes H.A. (1993) Biological classification of river water quality using neural networks, in *Applications of Artificial Intelligence in Engineering VIII,* eds. Rzevski G., Pastor J. and Adey R.A. Elsevier/CMP. 361-372.

Walley W.J., Hawkes H.A. and Boyd M. (1992) Application of Bayesian inference to river water quality surveillance, in *Applications of Artificial Intelligence in Engineering VII,* eds. Grierson D.E., Rzevski G. and Adey R.A. Elsevier/CMP. 1030-1047.

6 BIOGRAPHY

William Walley was formerly Senior Tutor in Civil Engineering at Aston University, where he taught water resources systems and researched in the field of computer modelling. In 1989 he teamed up with ecologist H. A. Hawkes to develop AI methods of river quality classification. He moved to Staffordshire University's School of Computing in 1993 to consolidate this work.

Sašo Džeroski is a researcher at the Jožef Stefan Institute, Ljubljana, having previously been a visiting researcher at the Turing Institute, Glasgow, and the Computer Science Department of the Katholieke University Leuvens, Belgium. He has researched into various areas of the theory and application of machine learning. He gained his BSc, MSc and PhD degrees from the Faculty of Electrical Engineering and Computer Science of Ljubljana University.