C. R. Biologies 335 (2012) 680–686

Contents lists available at SciVerse ScienceDirect

Comptes Rendus Biologies

www.sciencedirect.com

Ecology/Écologie

Support vector machines to model presence/absence of *Alburnus alburnus alborella* (Teleostea, Cyprinidae) in North-Western Italy: Comparison with other machine learning techniques

Tina Tirelli *, Marco Gamba, Daniela Pessani

Life Sciences and Systems Biology Department, Università degli Studi di Torino, Via Accademia Albertina 13, 10123 Torino, Italy

ARTICLE INFO

Article history:
Received 26 June 2012
Accepted after revision 9 September 2012
Available online 11 October 2012

Keywords:
Freshwater ecosystem
Decision trees
Artificial neural network
Support vector machines
Machine learning

ABSTRACT

Alburnus alburnus alborella is a fish species native to northern Italy. It has suffered a very sharp decrease in population over the last 20 years due to human impact. Therefore, it was selected for reintroduction projects. In this research project, support vector machines (SVM) were tested as possible tools for building reliable models of presence/absence of the species. A system of 198 sites located along the rivers of Piedmont in North-Western Italy was investigated. At each site, 19 physical-chemical and environmental variables were measured. We verified that performances did not improve after feature selection but, instead, they slightly decreased (from Correctly Classified Instances [CCI] = 84.34 and Cohen's k [k] = 0.69 to CCI = 82.81 and k = 0.66). However, feature selection is crucial in identifying the relevant features for the presence/absence of the species. We then compared SVMs performances with decision trees (DTs) and artificial neural networks (ANNs) built using the same dataset. SVMs outperformed DTs (CCI = 81.39 and k = 0.63) but not ANNs (CCI = 83.03 and k = 0.66), showing that SVMs and ANNs are the best performing models, proving that their application in freshwater management is more promising than traditional and other machine-learning techniques.

1. Introduction

Freshwaters, which are rapidly deteriorating all around the world, have been the focus of increasing attention [1–3]. This attention has inspired many ecological studies analyzing environmental and habitat factors affecting the distribution of freshwater organisms at different spatial scales. Worldwide, freshwater habitats, due to human disturbance, show an extinction rate of species that is predicted to be five times higher than that of terrestrial species and three times that of coastal marine mammals [4]. This stimulated the scientific community to develop practical tools for assessing running waters and species conditions ecologically and for suggesting management strategies.

We used in this research project *Alburnus alburnus alborella* (De Filippi, 1844), a subspecies native to northern Italy of the widespread European small cyprinid *Alburnus alburnus*. It has suffered a very sharp decrease in population over the last 20 years, although the causes are still unknown [5]. Because of this, it was selected for reintroduction projects [5]. Moreover, in Piedmont (NW Italy), a quite recent Regional Law (L.R. number 37 dated 29/12/06) lays down regulations for the management of running water fauna, habitat, and fishing, providing policies aimed at re-establishing consistent populations of native species and subspecies of freshwater fish fauna.

Among the promising tools that can help us solve such environmental challenges, like the loss of biodiversity, there are those that ecological informatics equips us with [6]. Ecological informatics can be seen as an interdisciplinary framework that uses advanced computational technology to study ecological processes and patterns on

* Corresponding author.
E-mail address: santina.tirelli@unito.it (T. Tirelli).

1631-0691/$ – see front matter
http://dx.doi.org/10.1016/j.crvi.2012.09.001

various levels of ecosystem complexity [7]. A rapidly growing area of ecological informatics is *machine learning* (ML), a tool that identifies structures in complex, nonlinear data and generates accurate predictive models.

Different applications of ML methods have been used in ecology [7–12], demonstrating that ML is a powerful alternative to traditional modelling approaches. ML methods consist of a range of approaches, including:

- artificial neural networks [13–18];
- classification and regression trees [17–25];
- fuzzy logic [26–28];
- genetic algorithms and programming [29];
- Bayesian belief networks [30];
- support vector machines [25,31–39].

All these methods are being used more and more. This is due to the fact that they can model the complex, nonlinear relationships that typify ecological data, without having to satisfy the restrictive assumptions of conventional, parametric approaches [40–43]. Further, they allow researchers to develop highly reliable models [7].

1.1. Support vector machines (SVMs)

SVMs consist of a new group of learning algorithms, originally developed by Vapnik [31]. They present a challenge for modellers because they are statistically based and because they guarantee performance in a theoretical way [44]. They are inductive modelling techniques inspired by some features of biological information processing. They are based on an algorithm that finds the maximum-margin hyperplane–i.e. the hyperplane showing the greatest separation between the classes. The instances at the minimum distances from the maximum-margin hyperplane constitute the support vectors. The maximum-margin hyperplane is defined exclusively by the set of support vectors. The support vectors are placed on the very edge of the class distributions inside the border region separating the classes. Therefore, they are elements that are critical for the training set. All the other training instances are irrelevant to the extent that they can be omitted without changing either the position or orientation of the hyperplane [36,45]. These support vectors are placed closest to the decision boundary. Therefore, the classifier uses extreme cases to separate the two classes from each other. For detailed descriptions of SVMs [31,46].

SVMs show several advantages over other ML techniques:

- unlikely reaching overfitting [31,46,47];
- producing results that are more competitive than those of the best current accessible-classification methods [25,39];
- yielding excellent generalisation performance while solving numerous nonlinear regression and time-series problems [25,31];
- requiring only a minimum of model tuning [25,32,48] and a small training dataset [36].

Especially this last point is very important in the ecological applications of the method because researchers can sample only a much smaller number of sites of extreme spectral response [36] avoiding conventional, more expensive sampling approaches [36].

On the other hand, they show some disadvantages: being computationally complex and slow. Even so, because of their many advantages, SVMs have been applied successfully to many tasks. Nevertheless, they have been applied to ecological predictions only in the last decade [22,32,33,35–39,44,48–51].

Because of their novelty and their potential usefulness in ecological applications, we decided to build models of *A. a. alborella* presence using the SVM approach.

The aims of the present research project are:

- to use SVMs to model the species' presence in Piedmont;
- to compare the SVM performance with the performances of decision trees (DTs) and artificial neural networks (ANNs) [18];
- to compare the performance of SVMs built under two different sets of circumstances:
  - without performing feature selection,
  - with the use of only those features that stem from a previous feature selection procedure.

This last comparison is made because feature reduction is an open question. In fact, some authors [25] deem feature reduction unnecessary for SVM classification, while others advocate feature reduction in order to make the classification and performance of the models more accurate [36,37,39]. Therefore, we aimed at understanding whether researchers do or do not need feature selection for the specific task of *A. a. alborella* modelling.

2. Materials and methods

2.1. Study area and data collection

The study system, which covers an area of 25,399 km^2, consisted of 198 sites located along the rivers of Piedmont in North-Western Italy. *A. a. alborella* was present at 110 of the sampling sites (55.56% of them).

Because generally data mining approaches are data driven, we chose variables generally accepted by experts according to their degree of importance for fish fauna [13,52–54]. We chose the following set of predictive variables: (1) altitude; (2) homogeneity in the width of the sampled tract (classes 0–5; the larger the widths of the sections examined, the larger the value); (3) amount of human impact (classes 0–5; the larger the impact, the larger the value); (4) amount of shade (classes 0–5; the larger the shade, the larger the value); (5) shelters for fish, visually assessed as the area consisting of undercut banks, macrophytae cover and debris jams (classes 0–5; the larger the cover, the larger the value); (6) percentage of bottom vegetation (algae and macrophytae) (classes 0–5; the larger the vegetation, the larger the value); (7–8–9) percentages of the sampled area with waterfalls classified according to their heights: (7) falls with heights > 1 m, (8) $0.5\ \text{m} \leq \text{high} \leq 1\ \text{m}$, (9) < 0.5 m; (10–11–12) percentages

of the sampled area classified according to water speed and depth: (10) riffles (areas of quite fast water with a broken-surface appearance), (11) pools (areas of slow, quite deep water with a smooth surface appearance), (12) flat reaches (areas with smooth constant depth and water speed), each reach surveyed and estimated visually; (13–18) percentages of the sampled area classified according to granulometry: (13) bedrock, (14) boulders and pebbles, (15) medium gravel (dimensions ≥ 1 cm), (16) little gravel (1 cm $<$ dimensions ≤ 2 mm), (17) sand (dimensions < 2 mm) and (18) silt; (19) pH.

The presence/absence data as well as the values of all 19 variables in each site were retrieved from the "Monitoraggio della fauna ittica in Piemonte" (Regione Piemonte, 2006).

A team of skilled ichthyologists collected data from spring to fall 2004. They used two types single-pass electrofishing: (1) a battery-powered electric fishing machine (AGK IG 200/2) operated at 150–300 V (the voltage varying according to the water conductivity); and (2) an internal-combustion-engine machine (EFKO FEG. 8000). The last one was used when the water was deeper than 1.5–2 m.

All 19 variables were submitted to the feature selection procedure. All inputs were comparable in terms of quality of data set over different sampling sites. Scales were standardised using z-scores – all river and habitat data were proportionally normalized between 0 and 1, minimum and maximum of all river and habitat-measured data ranged between 0.05 and 0.95.

For the classification phase we used SVMs. We built models including both the initial set of 19 features (indicated as 'non-feature selection' models, henceforth NFS models) and the subset features resulting from the feature selection (indicated as 'feature selection' models, FS models).

2.2. Feature selection phase

Feature selection is generally performed by searching the space of attribute subsets. This is done by combining an attribute-subset evaluator with a search method. In this study we used filter methods, which select features on the basis of measures of feature predictability and redundancy. A supervised filter is very flexible and allows various search and evaluation methods to be combined. In particular, we chose four supervised filter evaluators (χ^2, Information Gain, Gain Ratio, and Symmetrical Uncertainty) available in WEKA [45] with one search method (Ranker) to find the best feature set. For more details, see [18]. Moreover, feature selection was done by cross-validation (10-fold cross-validation) for each of the four methods. The algorithms used in each evaluator were exhaustively described by [45].

2.3. Model development

SVMs use Platt's sequential minimization algorithm (SMO) for training a support vector classifier [55–57]. This implementation replaces all missing values and transforms nominal attributes into binary ones. Platt's sequential minimization algorithm is also included in the machine-learning package WEKA (http://www.cs.waikato.ac.nz/ml/weka) [45].

We chose SMO because it is extremely easy to implement, often faster than other algorithms, and has better scaling properties. We applied the polynomial Kernel. We did not modify the default values of the parameter settings in the WEKA toolbox, except for the exponents of the polynomial Kernel. We tested different exponents from 1.0 to 5.0 to improve the performance of the SVM models [25]. The model with the best-performing exponent was chosen. Both for the FS subset as well as for NFS, we used k-fold cross-validation.

According to [16], the best k-value can be determined by building three different models: (1) models using a set of combinations of k between 3 and 10; (2) models using a set of combinations of k corresponding to the number of cases/2; and (3) models using a set of combinations of k corresponding to the number of cases $-$ 1. Therefore, we determined the optimal k value empirically by comparing the performances of different cross-validated SVMs using the Mann-Whitney U test.

Both for NFS SVMs and for FS SVMs, the model with the best-performing exponent and the best-performing k value was validated by using ten random subsets to estimate any eventual reliable error [58]. Random subsets were obtained using a custom syntax in IBM SPSS Statistics for Mac.

At this point, we ran the non-parametric Mann-Whitney U test to compare the performance of the NFS and FS models.

There are several ways to assess the performance of predictive models [59] each one with pros and cons. One of them is by calculating the percentage of sites where the presence/absence of the studied taxa is predicted correctly [60]. However, correctly classified instances (CCI) are affected by the frequency of occurrence of the test organism(s) being modelled [61–63]. To compensate, we used the following additional performance measures, namely: (1) model sensitivity (ability to predict species presence accurately); (2) model specificity (ability to predict species absence accurately); (3) Cohen's k coefficient [64]; and (4) the area under the receiver-operating-characteristic (ROC) curve. Cohen's k is a measure of the proportion of all possible cases of presence or absence that are predicted correctly after accounting for chance effects. Thus, Cohen's k interprets the predictive performance of the models better than CCI alone, being negligibly affected by prevalence [65–67]. Cohen's k gives a rather conservative estimate of prediction accuracy because it underestimates agreements due to chance [68]. According to literature [16,22,25,67], models with $k > 0.4$ and CCI $> 70\%$ are to be considered reliable.

Moreover, the authors of [69] suggest that different disciplines may show differences in k threshold values. Hence, they assess the following k values in a freshwater ecological context too, confirming the ranges suggested by [70], which are classified as 0.00–0.20, poor; 0.20–0.40, fair; 0.40–0.60, moderate; 0.60–0.80, substantial; and 0.80–1.00, almost perfect. Regarding the area under the ROC curve, a value of 0.7 indicates satisfactory discrimination, a value of

0.8 good discrimination and a value of 0.9 very good discrimination [71].

Starting from the results of a previous work [18], we used 10-fold cross-validated pruned DTs built using, among the 19 inputs, only the ones resulting from feature selection (see above). The reduction of features allows for a reduction in data-gathering and data-analyses costs, and also lead to better performance of DTs [18]. Similarly, we used ANNs built using features from the above mentioned feature selection. In models built after feature selection, the performance resulted significantly better than in those using all parameters [18]. We compared the performance of SVM models to those of ANNs showing the best-performing architecture in Tirelli and Pessani [18]: 9 input neurons, 6 hidden neurons, 1 output, and 10-fold cross-validated. DTs and ANNs were built using WEKA [45]. Model validation for DTs and ANNs was conducted following the same procedure as for SVM models. We ran the non-parametric Mann-Whitney U test to compare the performance of the SVM models built in the present study with the performances of DTs and ANNs [18]. This was done in order to ascertain the reliability of SVMs for fish fauna management.

3. Results and conclusions

3.1. Feature selection phase

We verified that all four of the methods converge on the selection of a unique core of relevant features, which are determined by applying the feature selection methods cited above as ranking methods for the overall set of features. The selected core is made up of the features present in the first 9 positions of the rankings. They are: (1) altitude; percentage of the sampled area showing (2) falls with heights < 0.5 m, (3) riffles and (4) flat reaches; the percentages of (5) bedrock, (6) boulders and pebbles, (7) little gravel; (8) sand and (9) pH.

We acknowledge that the selection of variables is not necessarily independent of the modeling approach (e.g. a variable that can be effective with ANNs may be ineffective with DTs). Nevertheless, we used the same set of variables to contrast the performances of different models.

3.2. Models performances

The best-performing models were obtained using an exponent of 1.7 for NFS models, of 1.9 for FS SVMs. The optimal k value was determined empirically by comparing the performances of different cross-validated SVM models using the Mann-Whitney U test. Among the different k-fold cross-validations that were tested for NFS models, the performances and reliability did not improve with $k > 99$. In fact, there were no statistical differences according to the results of the Mann-Whitney tests performed on the five parameters assessing the performances of the NFS models, between 99- and 197-fold cross-validated SVMs. Thus we used the 99-fold cross-validation to build our model. For FS models, the performances and reliability did not improve with $k > 10$. Thus we used the 10-fold cross-validation to build FS SVMs. Table 1 shows the mean performances of the different machine learning techniques (NFS SVMs, FS SVMs, ANNs, and DTs).

Table 1
Statistics of the performance of the machine learning models.

Model	Value	CCI	k	Sen	Spe	ROC
NFS SVMs	Mean	84.34	0.69	79.70	84.14	0.85
	St. Dev.	0.47	0.92	0.83	0.01	0.01
FS SVMs	Mean	82.81	0.66	79.94	86.04	0.83
	St. Dev.	1.18	1.17	1.83	0.02	0.01
ANNs	Mean	83.03	0.66	78.57	88.03	0.86
	St. Dev.	1.04	0.02	2.02	1.97	0.01
DTs	Mean	81.39	0.63	77.22	86.01	0.81
	St. Dev.	1.47	0.03	1.83	2.15	0.02

NFS SVMs: support vector machine models without feature selection; FS SVMs: support vector machine models built after selecting inputs using the four supervised-filter evaluators; ANNs: artificial neural network models; DTs: decision tree models; CCI: percentage of correctly classified instances; Sen: sensitivity; Spe: specificity; k: Cohen's k; ROC: area under the ROC curve; St. Dev.: standard deviation.

According to the performance parameters we considered, the presence/absence of *A. a. alborella* can be predicted reliably by SVMs. The average CCI, sensitivity and specificity either for NFS and FS SVMs was much higher than the threshold-limit value for considering a model reliable (= 70%). The same trend is followed for the mean value of k, which is much higher than the threshold. The values of the Cohen's k actually reveal substantially reliable models. In the end, also the area under the ROC curve shows, in both the models, very good discrimination.

Moreover, we conducted Mann-Whitney U tests to assess the statistical differences in the performances of the two types of SVM models: (1) the 10 repeated 99-fold cross-validated NFS models, (2) the 10 repeated 10-fold cross-validated FS models. The tests showed that the best predictions were obtained with NFS models, except for sensitivity (no statistical differences) and specificity (FS SVMs better performing, Table 2). Among the pull of 10 repeated 99-fold NFS SVMs, the best performing model had the following performances: CCI = 85.35%; Sen = 80.30%; Spe = 85.86; $k = 0.71$ and ROC = 0.86. Therefore, SVMs can make predictions very well.

FS SVMs performed slightly worse than NFS SVMs probably because feature selection can cause a loss of information about the impact of environmental and physical-chemical variables on *A. a. alborella* presence. However, as FS SVMs performances are much higher than the minimum standards of CCI = 70% and $k = 0.40$ and allow for identifying the most important features for *A. a. alborella* presence, they should be taken in consideration to build an ecologically-relevant model.

The fact that NFS SVMs show higher performances than FS SVMs is in contrast with what reported for other ML methods several times [17,18,39,72]. Learning in ANNs is sensitive to the inputs used. When choosing the appropriate features through pre-processing, models perform considerably better [72]. Without variable selection in ANNs, irrelevant information passes through the nodes, influences the connection weights, and affects the overall

Table 2
Performance comparisons between the different machine learning techniques.

Comparison	CCI	*k*	Sen	Spe	ROC
NFS SVMs vs FS SVMs	0.002*	0.002*	0.579	0.001*	0.004*
NFS SVMs vs ANNs	0.002*	0.003*	0.247	< 0.001*	0.015*
NFS SVMs vs DTs	< 0.001*	< 0.001*	0.002*	0.043*	0.001*
FS SVMs vs ANNs	0.481	0.631	0.123	0.052	< 0.001*
FS SVMs vs DTs	0.043*	0.035*	0.001*	0.631	0.052
ANNs vs DTs	0.007*	0.023*	0.190	0.063	< 0.001*

NFS SVMs: support vector machine models without feature selection; FS SVMs: support vector machine models built after selecting inputs using the four supervised-filter evaluators; ANNs: artificial neural network models; DTs: decision tree models; CCI: percentage of correctly classified instances; Sen: sensitivity; Spe: specificity; *k*: Cohen's *k*; ROC: area under the ROC curve; '*'denotes significant *P* values.

performance. Variable selection is fundamental because decreases ANN size, reduces computational costs, increases speed, and uses less data to estimate connection weights efficiently. Therefore it eliminates all but the most relevant attributes, reduces the number of input variables, and helps models predict better [67,72,73]. Moreover, the result of feature selection in SVMs absolutely confirms what Hoang and colleagues [25] reported: using feature selection methods in SVMs not necessarily increases the classification accuracy of the models significantly. But this fact is in disagreement with Sanchez-Hernandez and co-authors [36,37] and Favaro et al. [39]. We may explain this by the fact that SVM models are indeed more able to deal with a higher number of variables than other ML techniques. Even so, these models in certain circumstances still benefit from appropriate feature selection [39]. Our study shows that the choice of feature selection in SVM may not assure the best performing model but still guarantee high performances and may provide the researcher with ecologically relevant information.

We have endeavoured to determine the predictive model that performs the best because such a model can be used to manage properly *A. a. alborella*. Not all the modeling procedures we compared showed the same performance. NFS SVMs outperformed DTs (Table 2) except for specificity where DTs show better performances. FS SVM showed better performances when compared to DTs for CCI, Cohen's *k* and sensitivity, while no statistical differences have been found for specificity and area under the ROC curve (Table 2). ANNs (Table 2) outperformed DT models [18]. Different is the situation regarding the performance comparison between ANNs and SVMs. The Mann-Whitney tests showed that the best predictions were obtained with NFS SVMs for CCI and for Cohen's *k* and with ANNs for specificity and area under the ROC curve (Table 2). No difference was found for sensitivity. The comparison between FS SVMs and ANNs showed that ANNs have significantly higher values of the area under the ROC curve. All other performance parameters did not differed significantly.

Therefore, we can assert that both SVMs and ANNs are valuable and useful tools for predicting *A. a. alborella* presence/absence.

On the hand, our results are consistent with the analyses performed by Hoang and colleagues [25] to model the presence of macroinvertebrates in Vietnamese rivers. These authors showed better performances of SVMs over DTs, but unfortunately they did not compare the performances of SVMs with those of ANNs. On the other hand, the results of the present research are partially in disagreement with Favaro et al. [39], who reported SVMs as the best performing ML method to model the presence/absence of *Austropotamobius pallipes*.

In conclusion, *A. a. alborella*, being subjected to serious decline like most of the endemic freshwater fish species, needs researchers choose the best way to take on this decline by deeply understanding the relationships between species and their habitats. With this in mind, researchers can better plan management strategies. To improve our understanding of the ecological constraints to which the species is subjected, we should focus on the 9 inputs extracted during feature selection in FS SVMs. Both NFS and FS SVMs showed high performances: the former providing slightly higher results, the latter being more ecologically interpretable. In fact, a major disadvantage of the best performing NFS SVM models is the need to use all the 19 inputs, resulting in complex and less transparent models. In this way, the detection of general trends in the data is very difficult. Therefore, the ecological interpretation of the results may be a daunting task, as the NFS SVMs are unable to offer information about the habitat suitability for *A. a. alborella*. Feature selection is crucial to improve the transparency of the FS SVMs by reducing the number of inputs (from 19 to 9) and still allowing for a deeper look into the ecologically relevant parameters.

Among the input used in FS SVMs, altitude plays an important role because it is good integrator of the thermal conditions [18]. In this regard, our findings are in agreement with what reported for the fish community in New Zealand [74]. Flow velocity and substrate are crucial factors for the reproductive behaviour of this cyprinid, being essential for attaching eggs, therefore determining spawning habitat suitability [18]. Bedrock is a key factor for *A. a. alborella* because it provides shelters, which are fundamental when this species thrives. Tirelli et al. [72] referred the same situation for *Salmo marmoratus*. Bottom reaches of boulders and pebbles, and little gravel are also important for the presence of *A. a. alborella*, in agreement with Tirelli et al. [72]. Moreover, *A. a. alborella* often inhabits sandy bottom, while it is not found on silt bottoms to avoid physical alteration and infections (especially affecting gills) or because it interferes with its reproductive behaviour [18,72]. Prolonged exposure to fine suspended sediments can affect fish health and behaviour, causing changes in blood chemistry, gill- or skin-epithelia damage, and increasing the number of

infections. Moreover, it can result in higher mortality rates of both adults and embryos and delay the emergence of fry [18].

Therefore it is evident that it is extremely important to apply various ML techniques and contrast their performances to find out, when possible, the best performing model. Our own results show the advantages of contrasting various approaches. In fact our methods enabled us to predict *A. a. alborella* presence with reasonable accuracy and to identify key ecological factors. Had we used fewer approaches, we would have come up with a poorer model. Moreover it is to be underlined that it is not possible to answer to the one million dollar question: "What is the best performing ML method, the one that applies to all classification problems?" The best performing technique to solve one task, it is not necessary the best performing to solve a different one, as showed by *A. pallipes* and *A. a. alborella* cases. Therefore there is still a great deal of work to be done to improve the use of these kinds of approaches to ecology. However the use of multiple techniques will help both scientists and conservation professionals gaining more insight from their present and future data sets, both in terms of ecological relationships and of taxon-specific spatial distribution. This will contribute to improve management policies and thus conservation of biodiversity.

Disclosure of interest

The authors declare that they have no conflicts of interest concerning this article.

Acknowledgments

The authors would like to thank Regione Piemonte for granting access to "Monitoraggio della fauna ittica in Piemonte", Vincent Marsicano for the linguistic revision, and anonymous reviewers for their comments on a previous version of the manuscript.

References

[1] J.D. Allan, A.S. Flecker, Biodiversity conservation in running waters, Bioscience 43 (1993) 32–43.
[2] P.A. Matson, W.J. Parton, A.G. Power, M.J. Swift, Agricultural intensification and eco system properties, Science 277 (1997) 504–508.
[3] S.L. Postel, Entering an era of water scarcity: the challenges ahead, Ecol. Appl. 10 (2000) 941–948.
[4] A. Ricciardi, J.B. Rasmussen, Extinction rates of North American freshwater fauna, Conserv. Biol. 13 (1999) 1220–1222.
[5] C.M. Puzzi, A. Ippoliti, Sperimentazione di tecniche di reintroduzione dell'alborella (*Alburnus alburnus alborella*) negli ambienti lacustri della Provincia di Varese, Quaderni della Ricerca 36 (2004) 53.
[6] J.L. Green, A. Hastings, P. Arzberger, F.J. Ayala, K.L. Cottingham, K. Cuddington, F. Davis, J.A. Dunne, M.J. Fortin, L. Gerber, M. Neubert, Complexity in ecology and conservation: mathematical, statistical, and computational challenges, BioScience 55 (2005) 501–510.
[7] F. Recknagel, Ecological Informatics Understanding Ecology by Biologically-Inspired Computation, Springer-Verlag, Berlin and New York, 2003
[8] A.H. Fielding, Machine Learning Methods for Ecological Applications, Kluwer Academic Publishers, New York, 1999.
[9] F. Recknagel, Application of machine learning to ecological modelling, Ecol. Model. 146 (2001) 303–310, http://dx.doi.org/10.1016/S0304-3800(01)00313-1.
[10] J.B. Cushing, T. Wilson, Eco-informatics for decision makers advancing a research agenda, in: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences, Second International Workshop, DILS 2005, San Diego, CA, USA, Proceedings, Lecture Notes in Computer Science, 3615, Springer Verlag, Berlin, 2005, pp. 325–334.
[11] S. Ferrier, A. Guisan, Spatial modelling of biodiversity at the community level, J. Appl. Ecol. 43 (2006) 393–404.
[12] Y.S. Park, T.S. Chon, Biologically-inspired machine learning implemented to ecological informatics, Ecol. Model. 203 (2007) 1–7.
[13] S. Lek, A. Belaud, P. Baran, I. Dimopoulos, M. Delacoste, Role of some environmental variables in trout abundance models using neural networks, Aquat. Living Resour. 9 (1996) 23–29, http://dx.doi.org/10.1051/alr:1996004.
[14] H. Hoang, F. Recknagel, J. Marshall, J. Choy, Predictive modelling of macroinvertebrate assemblages for stream habitat assessments in Queensland (Australia), Ecol. Model. 146 (2001) 195–206.
[15] A.P. Dedecker, K. Van Melckebeke, P.L.M. Goethals, N. De Pauw, Development of migration models for macroinvertebrates in the Zwalm river basin (Flanders, Belgium) as tools for restoration management, Ecol. Model. 203 (2007) 72–86.
[16] P.L.M. Goethals, A.P. Dedecker, W. Gabriels, S. Lek, N. De Pauw, Applications of artificial neural networks predicting macroinvertebrates in freshwaters, Aquat. Ecol. 41 (2007) 491–508, http://dx.doi.org/10.1007/s10452-007-9093-3.
[17] T. Tirelli, D. Pessani, Use of decision tree and artificial neural network approaches to model presence/absence of *Telestes muticellus* in Piedmont (North-Western Italy), River Res. Appl. 24 (2009) 1001–1012.
[18] T. Tirelli, D. Pessani, Importance of feature selection in decision tree and artificial neural network ecological applications. *Alburnus alburnus alborella*: a practical example, Ecol. Inf. 6 (2011) 309–315, http://dx.doi.org/10.1016/j.ecoinf.2010.11.001.
[19] G. Death, K.E. Fabricius, Classification and regression trees: a powerful yet simple technique for ecological data analysis, Ecology 81 (2000) 3178–3192.
[20] S. Dzeroski, D. Demsar, J. Grbovic, Predicting chemical parameters of river water quality from bioindicator data, Appl. Intell. 13 (2000) 7–17.
[21] P.L.M. Goethals, S. Džeroski, P. Vanrolleghem, N. De Pauw, Prediction of Benthic Macroinvertebrate Taxa (Asellidae and Tubificidae) in Watercourses of Flanders by Means of Classification Trees, IWA 2nd World water congress, Berlin, 2001, pp. 5–6.
[22] E. Dakou, T. D'heygere, A.P. Dedecker, P.L.M. Goethals, M. Lazaridou-Dimitriadou, N. De Pauw, Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece), Aquat. Ecol. 41 (2007) 399–411, http://dx.doi.org/10.1007/s10452-006-9058-y.
[23] V. Lencioni, B. Maiolini, L. Marziali, S. Lek, B. Rossaro, Macroinvertebrate assemblages in glacial stream systems: a comparison of linear multivariatemethods with artificial neural networks, Ecol. Model. 203 (2007) 119–131.
[24] S. Pivard, D. Demsăr, J. Lecomte, M. Debeljak, S. Dzěroski, Characterizing the presence of oilseed rape feral populations on field margins using machine learning, Ecol. Model. 212 (2008) 147–154, http://dx.doi.org/10.1016/j.ecolmodel.2007.10.012.
[25] H. Hoang, K. Lock, A. Mouton, P.L.M. Goethals, Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam, Ecol. Inform. 5 (2010) 140–146, http://dx.doi.org/10.1051/kmae/2011037.
[26] A. Salski, C. Sperlbaum, A fuzzy logic approach to modeling in ecosystem research, in: B. Bouchon-Meunier, R.R. Yager, L.A. Zadeh (Eds.), Uncertainty in Knowledge Bases, 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU '90, Paris, France, July 2–6, 1990, Lecture Notes in Computer Science, 521, Springer-Verlag, Berlin, 1991, pp. 520–527.
[27] V. Adriaenssens, B. De Baets, P.L.M. Goethals, N. De Pauw, Fuzzy rule-based models for decision support in ecosystem management, Sci. Total Environ. 319 (2004) 1–12.
[28] A.M. Mouton, B. De Baets, P.L.M. Goethals, Knowledge-based versus data-driven fuzzy habitat suitability models for river management, Environ. Model. Softw. 24 (2009) 982–993.
[29] D.R.B. Stockwell, I.R. Noble, Induction of sets of rules from animal distribution data: a robust and informative method of analysis, Math. Comput. Simul. 33 (1992) 385–390.
[30] V. Adriaenssens, P.L.M. Goethals, J. Charles, N. De Pauw, Application of Bayesian belief networks for the prediction of macroinvertebrate taxa in rivers, Ann. Limnol. Int. J. Limnol. 40 (2004) 181–191.
[31] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[32] Q. Guo, M. Kellya, C.H. Graham, Support vector machines for predicting distribution of Sudden Oak Death in California, Ecol. Model. 182 (2005) 75–90.
[33] Q. Hu, C. Davis, Automatic plankton image recognition with co-occurrence matrices and Support Vector Machine, Mar. Ecol. Prog. Ser. 295 (2005) 21–31.

[34] J.M. Drake, C. Randin, A. Guisan, Modelling ecological niches with support vector machines, J. Appl. Ecol. 43 (2006) 424–432.
[35] Y. Shan, D. Paull, R.I. McKay, Machine learning of poorly predictable ecological data, Ecol. Model. 195 (2006) 129–138.
[36] C. Sanchez-Hernandez, D.S. Boyd, G.M. Foody, Mapping specific habitats from remotely sensed imagery: support vector machine and support vector data description based classification of coastal saltmarsh habitats, Ecol. Inform. 2 (2007) 83–88.
[37] C. Sanchez-Hernandez, D.S. Boyd, G.M. Foody, One-class classification for mapping a specific land-cover class: SVDD classification of Fenland, IEEE Trans. Geosci. Remote Sens. 45 (2007) 1061–1073.
[38] R. Ribeiro, L. Torgo, A comparative study on predicting algae blooms in Douro River, Portugal, Ecol. Model. 212 (2008) 86–91.
[39] L. Favaro, T. Tirelli, D. Pessani, Modelling habitat requirements of white-clawed crayfish (*Austropotamobius pallipes*) using support vector machines, Knowl. Managt. Aquatic Ecosyst. 401 (2011) 21.
[40] A. Guisan, N.E. Zimmermann, Predictive habitat distribution models in ecology, Ecol. Model. 135 (2000) 147–168.
[41] A.T. Peterson, D.A. Vieglais, Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem, Bioscience 51 (2001) 363–371.
[42] J.D. Olden, D.A. Jackson, A comparison of statistical approaches for modelling fish species distributions, Freshwater Biol. 47 (2002) 1976–1995.
[43] J. Elith, C.H. Graham, R.P. Anderson, M. Dudìk, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J.McC. Overton, A.T. Peterson, S.J. Phillips, K.S. Richardson, R. Scachetti-Pereira, R.E. Schapire, J. Soberón, S. Williams, M.S. Wisz, N.E. Zimmermann, Novel methods improve prediction of species' distributions from occurrence data, Ecography 29 (2006) 129–151.
[44] N. Cristianini, B. Scholkopf, Support vector machines and kernel methods–the new generation of learning machines, Ai Mag. 23 (2002) 31–41.
[45] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed, Morgan Kaufmann, San Francisco, 2005.
[46] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 3 (1998) 121–167.
[47] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, New York, John Wiley & Sons, 2001.
[48] D. Decoste, B. Scholkopf, Training invariant support vector machines, Mach. Learn. 46 (2002) 161–190.
[49] T. Joachims, Text categorization with support vector machines: learning with many relevant features, in: Proceedings of ECML-98, 10th European Conference on Machine Learning, Springer-Verlag, Berlin, (1998), pp. 137–142.
[50] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Natl. Acad. Sci. U S A 97 (2000) 262–267.
[51] C. Huang, L.S. Davis, J.R.C. Townshend, An assessment of support vector machines for land cover classification, Int. J. Remote Sens. 23 (2002) 725–749.
[52] S. Mastrorillo, S. Lek, F. Dauba, A. Belaud, The use of artificial neural networks to predict the presence of small-bodied fish in a river, Freshwater Biol. 38 (1997) 237–246, http://dx.doi.org/10.1046/j.1365-2427.1997.00209.x.
[53] J.D. Olden, D.A. Jackson, Fish–habitat relationships in lakes: gaining predictive and explanatory insight by using artificial neural networks, T. Am. Fish. Soc 130 (2001) 878–897.
[54] J.D. Olden, M.K. Joy, R. Death, Rediscovering the species in community-wide predictive modeling, Ecol. Appl. 16 (2006) 1449–1460.
[55] J.C. Platt, Fast training of support vector machines using sequential minimaloptimization, in: B. Schölkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods: Support Vector Machines, MIT Press, Cambridge, 1998, pp. 185–208.
[56] J.C. Platt, Using sparseness and analytic QP to speed training of support vector machines, in: M.S. Kearns, S.A. Solla, D.A. Cohn (Eds.), Advances in neural information processing systems, 11, MIT Press, Cambridge, 1999, pp. 557–563.
[57] S.S. Keerthi, S.K. Shevade, C. Bhattacharya, K.R.K. Murthy, Improvements to Platt's SMO algorithm for SVM classifier design, Neural Comput. 13 (2001) 637–649.
[58] J.D. Olden, J.J. Lawler, N.L. Poff, Machine learning methods without tears: a primer for ecologists, Q. Rev. Biol. 83 (2008) 171–193.
[59] A.M. Mouton, B. De Baets, P.L.M. Goethals, Ecological relevance of performance criteria for species distribution models, Ecol. Model. 221 (2010) 1995–2002.
[60] S. Manel, H.C. Williams, S.J. Ormerod, Evaluating presence/absence models in ecology: the need to account for prevalence, J. Appl. Ecol. 38 (2001) 921–931, http://dx.doi.org/10.1046/j.1365-2664.2001.00647.x.
[61] A.H. Fielding, J.F. Bell, A review of methods for the assessment of prediction errors in conservation presence/absence models, Environ. Conserv. 24 (1997) 38–49.
[62] S. Manel, J.M. Dias, S.T. Buckton, S.J. Ormerod, Alternative methods for predicting species distribution: an illustration with Himalayan river birds, J. Appl. Ecol. 36 (1999) 734–747.
[63] A.P. Dedecker, P.L.M. Goethals, W. Gabriëls, N. De Pauw, Comparison of artificial neural network (ANN) model development methods for prediction of macroinvertebrates communities in the Zwalm river basin in Flanders, Belgium, Scientific World J. 2 (2002) 96–104.
[64] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.
[65] A.P. Dedecker, P.L.M. Goethals, W. Gabriels, N. De Pauw, Optimisation of Artificial Neural Network (ANN) model design for prediction of macroinvertebrate communities in the Zwalm river basin (Flanders, Belgium), Ecol. Model. 174 (2004) 161–173.
[66] A.P. Dedecker, P.L.M. Goethals, N. De Pauw, Sensitivity and robustness of stream model based on artificial neural networks for the simulation of different management scenarios, in: S. Lek, M. Scardi, P.F.M. Verdonschot, J.P. Descy, Y.S. Park (Eds.), Modelling Community Structure in Freshwater Ecosystems, Springer-Verlag, Berlin, 2005 , pp. 133–146.
[67] T. D'heygere, P.L.M. Goethals, N. De Pauw, Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks, Ecol. Model. 195 (2006) 20–29.
[68] G.M. Foody, On the compensation for chance agreement in image classification accuracy assessment, Photogramm. Eng. Rem. S. 58 (1992) 1459–1460.
[69] W. Gabriels, P.L.M. Goethals, A.P. Dedecker, S. Lek, N. De Pauw, Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks, Aquat. Ecol. 41 (2007) 427–441.
[70] J.R. Landis, G.G. Koch, The measurements of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[71] D. Hosmer, S. Lemeshow, Applied Logistic Regression, second ed, John Wiley and Sons Inc, New York, 2000.
[72] T. Tirelli, L. Pozzi, D. Pessani, Use of different approaches to model presence/absence of *Salmo marmoratus* in Piedmont (Northwestern Italy), Ecol. Inform. 4 (2009) 234–242.
[73] T. D'heygere, P.L.M. Goethals, N. De Pauw, Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates, Ecol. Model. 160 (2003) 291–300.
[74] M.K. Joy, R.G. Death, Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks, Freshwater Biol. 49 (2004) 1036–1052.