Caspian J. Env. Sci. 2012, Vol. 10 No.2 pp. 237~246

CJES
Caspian Journal of Environmental Sciences

[Research]

Application of genetic algorithm (GA) to select input variables in support vector machine (SVM) for analyzing the occurrence of roach, *Rutilus rutilus*, in streams

R. Zarkami[1]*, R. Sadeghi Pasvisheh[2], P. Goethals[3]

1- Dept. of Environmental Science, Faculty of Natural Resources, University of Guilan, P.O. Box 1144, Sowmeh Sara, Guilan, Iran
2- Dept. of Plant Production, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, B–9000 Gent, Belgium
3- Dept. of Applied Ecology, Ghent University, J. Plateaustraat 22, B–9000 Gent, Belgium
** Corresponding author's E-mail: rzarkami2002@yahoo.co.uk*

(Received: Aug. 20-2011, Accepted: Feb. 05-2012)

ABSTRACT

Support vector machine (SVM) was used to analyze the occurrence of roach in Flemish stream basins (Belgium). Several habitat and physico–chemical variables were used as inputs for the model development. The biotic variable merely consisted of abundance data which was used for predicting presence/absence of roach. Genetic algorithm (GA) was combined with SVM in order to select the most important predictors for assessing the presence/absence of roach in the sampling sites. Before and after variable selection, the SVM were evaluated and compared by two predictive performances namely the percentage of Correctly Classified Instances (CCI %) and Cohen's kappa statistics (*k*). The obtained results showed that before variable selection, the SVM yielded a reliable performance but the prediction further improved after the combination of SVM with GA. According to the attribute weights, the habitat variables were more responsible than physico–chemical ones in assessing the presence/absence of fish in the streams. GA also presented that roach are more dependent on the habitat variables rather than on water quality ones. Though after variable selection the predictive performances increased, the attribute weights of SVM could be an alternative substitute for GA since all input variables can be evaluated in terms of their weights.

Keywords: *ecological modelling, genetic algorithm, roach (Rutilus rutilus), support vector machine*

INTRODUCTION

Predictive models are becoming more and more popular in various ecological studies in order to assess, monitor and manage natural resources. In fact, these models are mainly applied to predict and model the abundance and presence/absence of organisms in their habitat (Goethals *et al*., 2002; D'heygere *et al*., 2003, 2006; Dakou *et al*., 2007; Ambelu *et al*., 2010; Hoang *et al*., 2010; Zarkami *et al*., 2010). Models predicting presence/absence of organisms are very important in freshwater ecosystems (Jongman *et al*., 1995; Fielding and Bell., 1997). Predictive modelling is one of the most important steps in the development of a standard habitat assessment protocol (Parsons *et al*., 2004). Habitat use and the specific composition of communities are influenced by interactions between animals and their biotic and abiotic environment (Schoener, 1974; Begon *et al*., 1990).

Among the predictive models, the fish–habitat ones have an essential role in prioritizing surveys and monitoring programmes for fish populations (Jackson and Harvey, 1997). Habitat and spatial distribution of lake and river fish have long been studied (Copp, 1990; Rossier, 1995; Fischer and Eckmann, 1997; Brosse and Lek., 2000; Zarkami *et al*., 2010).

Since the last few years, various modelling techniques have been applied for the evaluation of running waters based on the distribution of organisms. Among these modelling techniques, SVM (Ambelu *et al*., 2010; Hoang *et al*., 2010), CT (Dzeroski *et al*.,

2000; Goethals *et al*., 2002; Dakou *et al*., 2007; Hoang *et al*., 2010; Ambelu *et al*., 2010; Zarkami *et al*., 2010), and fuzzy logic (Adriaenssens *et al*., 2004; Mouton *et al*., 2009) have shown to be very powerful methods when analyzing the habitat suitability of organisms.

SVM implements Platt's sequential minimal optimization algorithm (Platt, 1998) for training a support vector classifier (Keerthi *et al*., 2001). Multi–class problems are solved using pair-wise classification (Witten *et al*., 2011). SVM consists of input and output layers connected with weight vectors. Since the last few years till now, researchers have become more interested in applying SVM (Vapnik, 1995; Burges, 1998; Keerthi *et al*., 2001; Ambelu *et al*., 2010) because it gives excellent generalization performance on a wide range of problems (Keerthi *et al*., 2001). It is often much faster and has better scaling properties (Keerthi *et al*., 2001). SVM makes very competitive results with the best accessible classification methods and needs only the smallest amount of model tuning (Decoste and Scholkopf, 2002; Guo *et al*., 2005).

The present study aimed to develop models that could predict the occurrence of roach in the Flemish streams (Belgium) using SVM in combination with genetic algorithms (GA) and then to compare the predictive performance of SVM before and after GA. These models would allow the selection of the most important variables for river restoration.

MATERIALS AND METHODS

Study area

Several river basins were monitored during the study period. The Scheldt is the main and biggest river basin in Flanders (Belgium) with a total surface of around 19150 km². For management plans, the Scheldt is divided into several sub basins. The Yzer, the second important river basin located in the coastal region, has a surface area of about 734 km². The surface area of Demer basin is around 2280 km². The lower part of the Scheldt and the Yzer are estuarine with tidal effects. Most running waters in Flanders are as the lowland brook type. In contrast to running water, the standing waters are contaminated. Dender has an irregular flow regime so that more than 90 % of the surface water of the Dender consists of rainwater. During summer algal blooms occur very often because of low flow velocity. Most river basins are impacted by domestic, agricultural and industrial activities causing various pollution types in Dender. Fish population are faced with problems due to artificial embankment, dams and weirs. In addition to this, their migration paths encounter trouble. Fig.1 illustrates the main river basin for sampling.

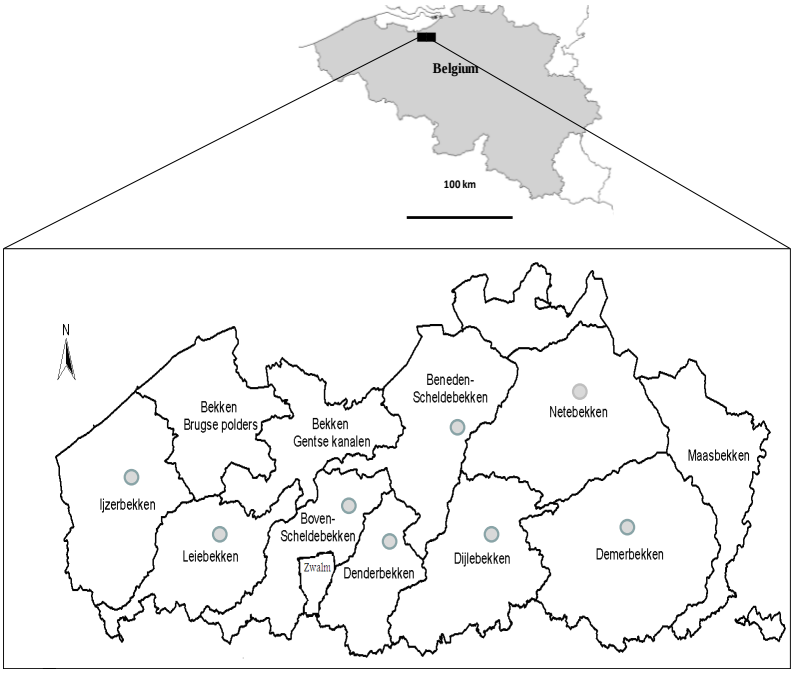


Fig 1. Location of stream basins for fish sampling.

Data sources and collection

Based on Fig.1, the main stream basins were considered to take the biotic and abiotic sampling into account which resulted in nearly 180 instances. At each site, various abiotic variables (habitat and physico–chemical measurements) were examined. Samples were taken on each month during a few years investigation so there was at least one electro fishing event for each month. The abiotic variables were used as inputs but the biotic one as output for the development of SVM. The biotic variable merely consisted of the abundance data which served for analyzing the presence/absence of fish.

Weka toolbox software (Witten *et al*., 2011, version 3.4.18, 1999–2010) was used to develop the SVM. The observed values of fish were considered 50 % in the sampling campaigns. A standardized electro fishing method was used to take fish samples from the streams. This device was equipped with a 5 kW generator and an adjustable output voltage of 300/500V and a pulse frequency of 480 Hz. This method was conducted over stream (or river) lengths of at least 20 times the stream (or river) width. This electric–fishing had two hand–held anodes except when the river was smaller than 1 m.

Co–ordinate (X and Y) and site code were imported to ArcView (version 3.2a) to produce a geographical map for presenting the presence/absence of roach in the monitored sites (X and Y were respectively indicated for the directions from West to east and from North to south). The habitat and physico–chemical stream characteristics were collectively determined in the field and laboratory based on standardized and quality controlled methods.

Model development and optimization

The first step was to develop SVM using all input variables. Then SVM was combined with GA for the selection of the most explanatory input variables for the target fish. For the training and validation of SVM, a three–fold cross–validation (as indicated as supplied test set in Weka toolbox) was used to get a reliable estimate of the error of each model (Kohavi, 1995; Dakou *et al*., 2007; Witten *et al*., 2011). Cross–validation is a statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis (Witten *et al*., 2011).

Model performances are key components of the model training and validation procedures. This can be an important starting point to assess the quality of the models. Various model evaluations have been proposed in literature in order to meet this purpose of which the percentage of correctly classified instances (CCI %) and kappa statistics (k) are very popular ones when assessing the presence/absence of organisms (Goethals *et al*., 2007; Witten *et al*., 2011).

Attribute selection and optimization using GA

GA is a general purpose search algorithms inspired by Charles Darwin's principle of "survival of the fittest" to solve complex optimisation problems (Holland, 1975; Goldberg, 1989; Vose, 1999). The basic idea is to maintain a population of chromosomes.

In the present study, to allow input variable selection, the SVM was combined with GA. The attribute evaluator was based on a wrapper subset evaluator function with a full training set. Here, also a three–fold cross–validation was considered in order to estimate the accuracy of the learning scheme for a set of attributes. This was obtained by trial and error (in Weka toolbar, the default value is set on 5 folds). The Wrapper Subset Evaluator function ('weka.attributeSelection. WrapperSubsetEval') evaluates attribute sets by using a learning scheme. In the wrapper method, the variable selection algorithm functions as a wrapper for SVM. After the new variables were selected by GA, the model performance criteria were compared with and without GA. The number of genes in chromosomes was equal to the number of input variables. The initial population consisted of 20 chromosomes evolved through a maximum of 20 generations. Crossover was set at a probability of 60 %. Other settings were set as default as follows, the probability of mutation, 3.3 %, the number of fold, 3 and significant level, 0.1.

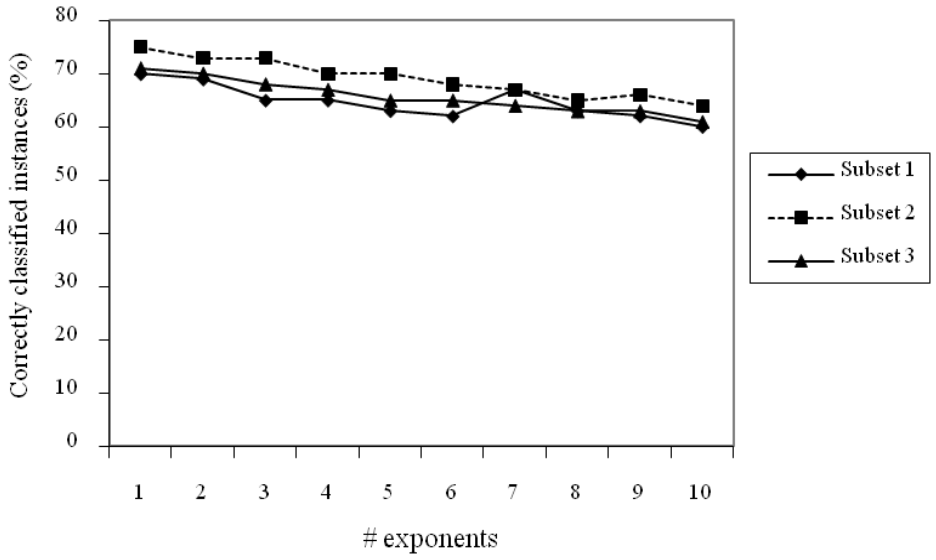
RESULTS

Model development and validation

Fig. 2 shows the outcomes of 3 subsets (based on a three–fold cross–validation) to assess the predictive performances of the SVM. Here, the dataset was split into three equal partitions and each in turn was used for validation, while the remainder was used for training (Witten *et al*., 2011). Therefore, from the total instances recorded, two–third of total dataset was used for training and one–third for validation. This method was applied so as to predict the error rate of a learning algorithm.

Most of the default settings were used in SVM except for the exponent of the polynomial kernel. To find the optimal predictive performances in order to forecast the presence/absence of roach, the SVM was optimized based on the application of different exponents from one to ten. Based on the obtained results, the highest performance was obtained in the exponent one (Fig. 2) so that the means of CCI and *k* met the threshold value so as to have a trustworthy prediction (CCI > 70 % and *k* > 0.40). Then, this best performing exponent was used to choose the most important predictors for fish in stream basins (Fig.3).

a



b

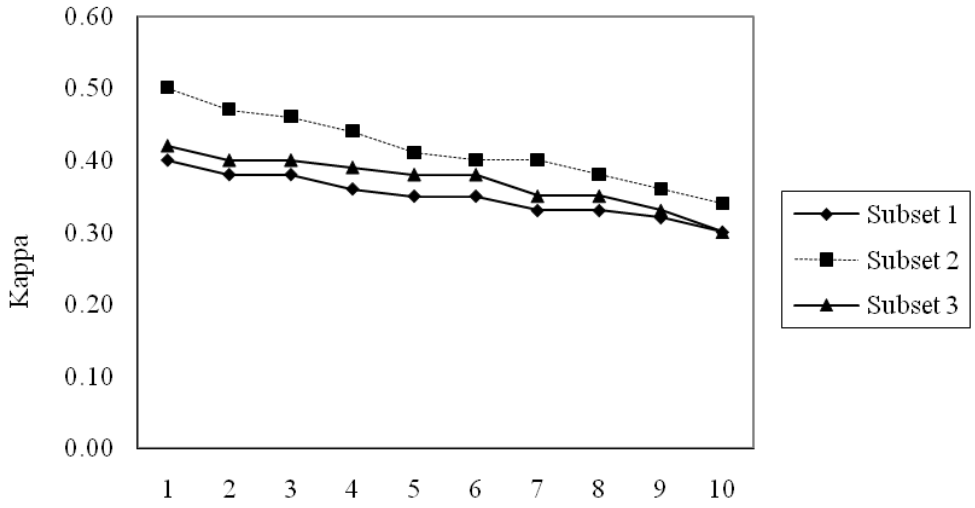


Fig 2. The predictive performances of SVM based on correctly classified instances (%) (a) and Kappa statistics (b) with applying different exponents.

Attribute weights

Fig. 3 illustrates the importance of attribute weights in relation to the most important input variables in the SVM when judging the presence/absence of roach. Based on the outcomes, the attribute weights were different for each input variable. A variable was considered as an important predictor when it had an absolute weight value greater than 0.5. The attribute weights of SVM are able to show the contribution of the all input variables to the prediction but the extent of weights should determine the importance of each predictor. The most important variables consisted of distance from the source, width, depth and slope. The water quality variables like EC, total phosphate, DO and nitrate (NO_3^-) were to a lesser extent important. Some variables (e.g. pH and flow velocity) were moreover presented by SVM but they had less effect on fish.

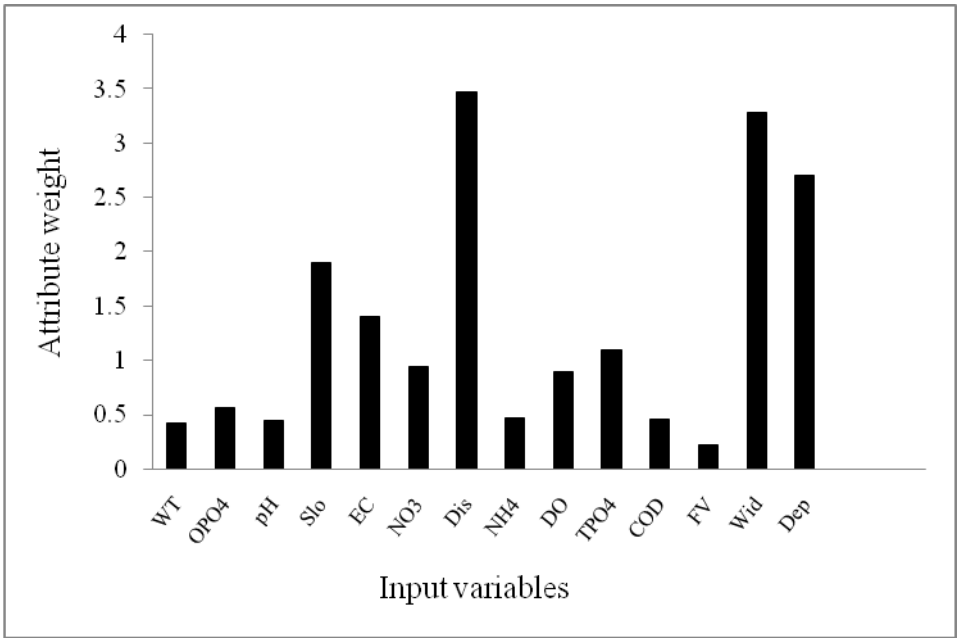


Fig 3. The selected input variables based on attribute weight in SVM (W.T.°C: water temperature; Dep: depth; Slo: slope; EC: electric conductivity; DO: dissolved oxygen; Wid: width; Dis: distance from the source; FV: flow velocity).

Variable selection using GA

Table 1 compares the predictive results of a three–fold cross–validation before and after variable selection. The obtained results showed that before variable selection, the values of CCI and *k* (mean ± Stdv) were reliable (p<0.001; CCI=75.7 % ± 2.1 and k=0.44 ± 0.05) but after GA, the prediction outcomes increased (P<0.001; CCI=79.7 % ± 2.08 and k=0.60 ± 0.03).

Table 1. Comparison of predictive results of SVM (with exponent=1) before and after GA

	Before variable selection		After variable selection	
Folds	CCI (%)	*k*	CCI (%)	*k*
Fold 1	75	0.50	78	0.57
Fold 2	78	0.57	82	0.63
Fold 3	74	0.49	79	0.60
	(75.70 ± 2.08)*	(0.52 ± 0.04)*	(79.70 ± 2.08)*	(0.60 ± 0.03)*

*(mean ± Stdv)

According to Table 2, the most important selected variables by GA were distance from the source, width and depth (each repeated in 3 times) and slope (repeated in 2 times). The variables that were repeated only one time were pH, water temperature, electric conductivity, nitrate and COD. Other variables (e.g. flow velocity and etc) were never recognized as important predictors by GA.

Table 2. Main variables selected by GA with a three–fold cross–validation in SVM (EC: electric conductivity, DO: dissolved oxygen, W.T.°C: water temperature)

	SS1	SS2	SS3
#selected variables	Slope, Distance, EC, Width, COD, Depth	Distance, Width, Depth, NO_3	Depth, Slope, Distance, W.T, Width
Three times	Distance, Width, Depth		
Two times	Slope		
One time	pH, COD, W.T, NO_3, EC		

DISCUSSION

There is a relationship between the number of instances as training and the number of irrelevant attributes. This means that the number of training instances for producing a suitable performance increases exponentially with number of irrelevant attributes (Witten *et al*., 2011). On the basis of this, the number of adequate instances for training and validation sets of SVM in predicting the fish was obtained by trial and error.

Measuring the predictive model performances (like CCI % and *k* applied in the present work) frequently entails calculating the percentage of the sites for which presence/absence of organisms is correctly predicted (Manel *et al*., 2001). Many authors (Fielding and Bell., 1997; Manel *et al*., 1999; Goethals and De Pauw, 2001; Dedecker *et al*., 2002; D'heygere *et al*., 2003; Dakou *et al*., 2006; Hoang *et al*., 2010; Zarkami *et al*., 2010) explored that frequency of occurrence can affect the percentage of CCI. When the organisms are very common or extremely rare, the number of correctly classified instances is very high during the validation process, but this can mainly be explained by the high reliability to make a good prediction even without making use of information from the data. Therefore, this study tried to examine another evaluation index called *k* (Cohen's kappa) to obtain a trustable result (Goethals, 2005). Since the frequency of roach occurrence was considered 50 % in all sampling campaigns, a logical relationship was expected to be obtained between CCI % and *k*. So by using a three-fold cross–validation, SVM showed reasonable outcomes between predicted and observed values for the target fish in the sampling sites.

Ecological modelling dealing with habitat requirements and the prediction of organisms are considered as a useful method to support decision–making in river restoration management (Goethals and De Pauw, 2001; Hoang *et al*., 2010; Zarkami *et al*., 2010). Therefore, for the management goals, it is very important to make a decision with regard to the selection of the most explanatory predictors for aquatic organisms (Goethals, 2005; Ambelu *et al*., 2010; Hoang *et al*., 2010; Zarkami *et al*., 2010). That is why a search method (GA) was used and combined with SVM for the selection of the major input variables for fish. Hoang *et al*. (2010), for instance, combined SVM and classification tree (CT) with GA so as to predict presence/absence of macroinvertebrates in the Du River (Northern Vietnam). Ambelu *et al*. (2010) conducted almost the same study in Gilgel Gibe watershed in Ethiopia. According to the authors, SVM yielded excellent performances than CT. D'heygere *et al*. (2006), moreover, developed GA in combination with ANN and CT, predicting the presence or absence of benthic macroinvertebrate taxa in unnavigable watercourses in Flanders (Belgium).

According to the attribute weights, however, the all input variables contributed to the prediction, a high relationship was noticed between the occurrence of fish and habitat variables in particular the distance from source and width were among the main predictors. Brosse and Lek (2000) showed that the most important variables influencing the 0+roach distribution was distance from the bank, depth, local slope of the bottom, percentage of mud and flooded vegetation cover. The importance of all structural–habitat variables for fish was also confirmed after the variable selection by GA. The depth is an important predictor for the habitat use of roach (Brabrand and Faafeng, 1994; Garner, 1995). Depth forms an essential feature in 0+roach habitat preference considering the needs for shelters against predation. Roach avoid the deep water and steeply sloping parts because these areas are usually occupied

by some top predators e.g. perch (*Perca fluviatilis* L.), pike (*Esox lucius*) (Brabrand and Faafeng, 1994; Eklöv, 1997).

In contrast to the habitat variables, the contribution of water quality ones to the prediction was negligible. Nevertheless, the selected variables might not be considered the only important ones in such a forecasting model. However, SVM is less affected by missing data (Witten *et al*., 2011), multiple collinearity between variables (high correlation) might cause a possible noise in data driven model. This would confuse the predictive models in selecting both variables for the given organisms. Another possible reason could be that roach are dominant fish species under eutrophic conditions (Persson, 1983). They are able to survive a wide range of environmental condition.

Some valuable inputs variables (e.g. the percent of vegetation cover) were eliminated from the dataset due to too missing values. For instance, roach are strongly associated with aquatic vegetation (Garner, 1995; Rossier *et al*., 1996). Some variables such as flow velocity, on the other hand, were introduced to the model but they were not recognized as important predictors. The problem was that flow velocity was not measured regularly during the study period so that there was insufficient information about this variable. Habitat use by roach varies in lakes and streams, where current velocity effectively influences their habitat (Moyle and Baltz, 1985; Copp, 1992). As a result, the reliability of model might be further improved with monitoring more relevant variables in the standard monitoring network.

CONCLUSIONS

The present study aimed to examine the occurrence of roach (based on presence/absence information) using SVM technique. In addition to this, a search algorithm method (GA) was combined with the SVM in order to select the most important predictors for roach. The comparison of the predictive performances (with and without optimization of GA) revealed that the developed model was reliable but the reliability of SVM became more prominent after variable selection. In spite of this, attribute weights of SVM could be an alternative to GA to select input variables since all attributes can be measured based on their weights. The information obtained in such a way could be useful for river management and restoration purposes. It can be concluded that focusing on the structural habitat of streams would be the main priority for stream management. Besides, the minimization of nutrient inputs and organic waste discharges would significantly improve the ecological quality in the streams.

ACKNOWLEDGEMENT

Data were provided by the laboratory of aquatic ecology and environmental toxicology (Ghent University, Belgium).

REFERENCES

Adriaenssens, V., De Baets, B., Goethals, P.L.M. and De Pauw, N. (2004). Fuzzy rule–based models for decision support in ecosystem management. *Science of the Total Environment.* 319, 1–12.

Ambelu, A., Lock, K. and Goethals, P. (2010). Comparison of modelling techniques to predict macroinvertebrate community composition in rivers of Ethiopia. *Ecological informatic*. 5, 147–152.

Begon, M., Harper, J.L. and Townsend, C.R. (1996). Ecology, Individuals, Population, and Communities, 3rd edn. Blackwell Science, Oxford.

Brabrand, A. and Faafeng, B. (1994). Habitat shift in roach, *Rutilus rutilus* induced by the introduction of pike-perch, *Stizostedion lucioperca*. *Limnologie*. 25, 21–23.

Brosse, S. and Lek, S. (2000). Modelling roach (*Rutilus rutilus*) microhabitat using linear and nonlinear techniques. *Freshwater Biology.* 44, 34–41.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*. 2, 121–167.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20, 37–46.

Copp, G. H. (1990). Shifts in the microhabitat of larval and juvenile the roach, Rutilus rutilus L. in a floodplain channel. *Journal of Fish Biology*. 36, 683–692.

Copp, G. H. (1992). An empirical model for predicting microhabitat of 0+ juvenile fishes in a lowland river catchment. *Oecologia*. 91, 338–345.

Dakou, E., D'heygere, T., Dedecker, A. P., Goethals, P.L.M., Lazaridou–Dimitriadou, M. and De Pauw, N. (2007). Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). *Aquatic Ecology*. 41, 399–411.

Dakou, E., Goethals, P.L.M., D'heygere, T., Dedecker, A.P., Gabriels, W. and De Pauw, N. (2006). Development of artificial neural network models predicting macroinvertebrate taxa in the river Axios (Northern Greece). *Japanese Journal of Limnology*. 15, 10–17.

Decoste, D. and Scholkopf, B. (2002). Training invariant support vector machines. *Machine Learning*. 46, 161–190.

Dedecker, A. P., Goethas, P.L.M., Gabriels, W. and De Pauw, N. (2002). Comparison of Artificial Neural Network (ANN) model developments methods for prediction of macroinvertebrates communities in the Zwalm river basin in Flanders, Belgium. *The Scientific World Journal*. 2, 96–104.

D'heygere, T., Goethals, P. L. M. and De Pauw, N. (2006). Genetic algorithms for optimization of predictive ecosystems models based on decision trees and neural networks. *Ecological Modelling*. 195, 20–29.

D'heygere, T., Goethals, P. L. M. and De Pauw, N. (2003). Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinverteberates. *Ecological Modelling*. 160, 291–300.

Dzeroski, S., Demsar, D. and Grbovic, J. (2000). Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence*. 13, 7–17.

Eklöv, P. (1997). Effects of habitat complexity and prey abundance on the spatial and temporal distributions of perch (*Perca fluviatilis*) and pike (*Esox lucius*). *Canadian Journal of Fisheries and Aquatic Sciences*. 54, 1520–1531.

Fielding, A. H. and Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*. 24, 38–49.

Fischer, P. and Eckmann, R. (1997). Spatial distribution of littoral fish species in a large European lake, Lake Constance, Germany. *Archiv für Hydrobiologie*.140, 91–116.

Garner, P. (1995). Suitability indices for juvenile 0+roach, *Rutilus rutilus* (L.) using point abundance sampling data. *Regulated Rivers: Research and Management (SAUS)*. 10, 99–104.

Goethals, P.L.M. and De Pauw, N. (2001). Development of a concept for integrated ecological river assessment in Flanders, Belgium. *Journal of Limnology*. 60, 7–16.

Goethals, P. L. M. (2005). Data driven development of predictive ecological models for benthic macroinvertebrates in rivers. PhD thesis. University of Ghent. 377 pp.

Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S. and De Pauw, N. (2007). Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquatic Ecology*. 41, 491–508.

Goethals, P.L.M., Dedcker, A., Gabriels, W. and De Pauw, N. (2002). Development and application of predictive river ecosystem models based on classification trees and artificial neural networks. Ecological informatics, Understanding ecology by biologically inspired computation. (ed. Recknagel), Springer, Berlin, 432 pp.

Goethals, P.L.M. and De Pauw, N. (2001). Development of a concept for integrated ecological river assessment in Flanders, Belgium. *Journal of Limnology*. 60, 7–16.

Goldberg, D. E. (1989). Genetic Algorithms in Search Optimization, and Machine Learning, Addison–Wesley, Reading, MA.

Guo, Q., Kellya, M. and Graham, C.H. (2005). Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling*.182, 75–90.

Hoang, T.H., Lock, K., Mouton, A. and Goethals, P. L.M. (2010). Application of classification trees and support vector machines to model the presence of

macroinvertebrates in rivers in Vietnam. *Ecological Informatic*. 5, 140–146.

Holland, J.H. (1975). Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.

Jackson, D.A. and Harvey, H.H. (1997). Qualitative and quantitative sampling of lake fish communities. *Canadian Journal of Fisheries and Aquatic Sciences.* 54, 2807–2813.

Jongman, R.H.G., Ter Braak, C. J. F. and Van Tongeren, O.F.R. (1995). Data Analysis in Community and Landscape Ecology, 2nd ed. Cambridge University Press, Cambridge, p. 299. *Journal of Futures Markets.* 15, 953–970.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C. and Murthy, K.R.K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation.* 13, 637–649.

Kohavi, R. (1995). A study of cross–validation and bootstrap for accuracy estimation and model selection. In: Lavrac, M., Wrobel, S. (Eds.), Proceedings of the International Joint Conference on Artificial Intelligence. Pp. 1137–1143.

Manel, S., Williams, H.C. and Ormerod, S.J. (2001). Evaluating presence–absence models in ecology, the need to account for prevalence. *Journal of Applied Ecology*.38, 921–931.

Manel, S., Dias, J.M., Buckton, S.T. and Ormerod, S. J. (1999). Alternatives methods for predicting species distribution, an illustration with Hialayan river birds. *Journal of Applied Ecology*.36,734–747.

Mouton, A.M., De Baets, B. and Goethals, P.L.M. (2009). Knowledge–based versus data–driven fuzzy habitat suitability models for river management. *Environmental modelling software*. 24, 982–993.

Moyle, P.B. and Baltz, D.M. (1985). Microhabitat use by an assemblage of California stream fishes, developing criteria for in–stream flow determinations. *Transactions of the American Fisheries Society*. 114, 695–704.

Parsons, M., Thoms, M.C. and Horris, R.H. (2004). Development of a standard approach to river habitat assessment in Australia. *Environmental Monitoring and Assessment.* 98, 109–130.

Persson, L. (1983). Effects of intraspecific and interspecific competition on dynamics and size structure of a perch, Perca fluviatilis and a roach, Rutilus rutilus population. *Oikos*. 41, 26–32.

Platt, J. (1998). "Fast Training of Support Vector Machines using Sequential Minimal Optimization". Advances in Kernel Methods–Support Vector Learning, eds: Schoelkopf, B., Burges, C. and Smola, A., MIT Press.

Rossier, O., Castella, E. and Lachavanne, J.B. (1996). Influence of submerged aquatic vegetation on size class distribution of perch (*Perca fluviatilis*) and roach (*Rutilus rutilus*) in the littoral zone of Lake Geneva (Switzerland). *Aquatic Sciences.* 58, 1–14.

Rossier, O. (1995). Spatial and temporal separation of littoral zone fishes of Lake Geneva (Switzerland–France). *Hydrobiologia*. 300/301,321– 327.

Schoener, T. (1974). Resource partitioning in ecological communities. *Science*. 185, 27– 39.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer–Verlag, New York.

Vose, M.D. (1999). Random heuristic search. *Theoretical Computer Science.* 229, 103–142.

Witten, I.H., Frank, E. and Hall, M.A. 2011. Data Mining, Practical Machine Learning Tools and Techniques. *Morgan Kaufmann,* San Francisco, 3rd ed. 629 pp.

Zarkami, R., Goethals, P. and De Pauw, N. (2010). Use of classification tree methods to study the habitat requirements of tench (*Tinca tinca*) (L., 1758). *Caspian journal of environmental science*. 8, 55–63.

کاربرد ژنتیک الگوریتم در انتخاب متغیرهای ورودی در ماشین بردار پشتیبان به منظور تجزیه و تحلیل میزان وقوع ماهی کلمه در رود خانه ها

ر. زرکامی*، ر. صادقی پس‌ویشه، پ. گوتالس

(تاریخ دریافت:۹۰/۵/۳۰ - تاریخ پذیرش:۹۰/۱۱/۱۷)

چکیده

در این کار تحقیقی، مدل **ماشین بردار پشتیبان** (SVM) برای بررسی میزان وقوع کلمه در رودخانه‌های فلامان در بلژیک مورد استفاده قرار گرفته است.برای انجام این کار چندین پارامتر فیزیکی-شیمیایی و ساختاری محیط (که اصطلاحا به فاکتورهای محیطی معروف هستند) در رود خانه‌ها اندازه گیری شده اند. این فاکتور ها به عنوان درون داد (input) برای توسعه مدل در نظر گرفته شده اند. تنها متغیر زیستی برای مطالعه میزان پراکنش ماهی اندازه گیری میزان فراوانی کلمه (*Rutilus rutilus*) بوده است.در واقع این فاکتور برای پیش بینی حضور و یا عدم حضور ماهی (نقش برون داد output در مدل) به کار گرفته شده است. به منظور انتخاب مناسب‌ترین فاکتورها در پیش بینی میزان پراکنش ماهی یک تکنیکی به نام ژنتیک الگوریتم (Genetic algorithm) با ساپورت وکتور ماشین ترکیب شده است. در صد قابلیت اعتماد مدل‌ها چه قبل و چه بعد از انتخاب متغیر ها، با دو تا از متداول‌ترین شاخص‌های آماری سنجیده شده است: ۱- شاخص تعداد داده‌های که به درستی کلاسه بندی شده اند (CCI٪) ۲- کاپای کوهنی (Cohen kappa). بر اساس نتایج حاصله، قبل از به کار گیری ژنتیک الگوریتم ، قابلیت اعتماد بالایی در ساپورت وکتور ماشین حاصل شد. اما بعد از ان که ساپورت وکتور ماشین با ژنتیک الگوریتم ترکیب شد در صد قابلیت اعتماد مدل بسیار بالاتر رفت. بر اساس معیار سنجش وزنی متغیر ها، متغیر های ساختاری-زیستگاهی بیش از متغیرهای فیزیکی-شیمیایی در پیش بینی میزان حضور و عدم حضور نقش داشتند. این متغیرها نیز بعد از به کار گیری ژنتیک الگوریتم نیز تاٌیید شده است. اگر چه بعد از انتخاب متغیرها توسط ژنتیک الگوریتم در صد قابلیت اعتماد مدل‌ها افزایش یافت با این وجود سنجش وزنی متغیرها می‌تواند یک جایگزین مناسبی برای ژنتیک الگوریتم باشد چون تمام متغیرها را می‌توان بر اساس وزن انها کلاسه بندی کرد در صورتی که در ژنتیک الگوریتم فقط تعداد مهم بودن و یا نبودن فاکتورها بررسی می‌گردد.

*مولف مسئول