Studi Trent. Sci. Nat., 87 (2010): 255-256 ISSN 2035-7699

Nota breve - Short note

Use of two data mining techniques to model presence/absence of *Telestes muticellus*, in Piedmont (North-Western Italy)

Tina TIRELLI* & Daniela PESSANI

Dipartimento di Biologia Animale e dell'Uomo, Università degli Studi di Torino, Via Accademia Albertina 13, 10123 Torino, Italy
* E-mail dell'Autore per la corrispondenza: *santina.tirelli@unito.it*

RIASSUNTO - *Utilizzo di alberi decisionali e reti neurali artificiali per costruire modelli predittivi della presenza/assenza di* Telestes muticellus *in Piemonte* - Esistono diverse tecniche ben conosciute di machine-learning altamente adatte a costruire modelli predittivi. E' quindi estremamente utile l'applicazione di tali strumenti per valutare lo stato ecologico dei fiumi e delle specie in essi presenti. Abbiamo costruito modelli predittivi di presenza/assenza di *Telestes muticellus*, Cyprinidae minacciato di estinzione, utilizzando reti neurali artificiali ed alberi decisionali. Questi ultimi hanno performance soddisfacenti e predizioni attendibili; il post-pruning ne riduce la complessità, aumentandone la trasparenza. Le reti neurali hanno performance migliori rispetto agli alberi e dimostrano il loro elevato potenziale per preservare l'ecosistema delle acque interne e prendere decisioni inerenti la gestione di specie minacciate.

Key words: decision tree, artificial neural network, ecological modelling, pruning, Piedmont, species prediction
Parole chiave: alberi decisionali, reti neurali, modelli ecologici, ottimizzazione con potatura, Piemonte, predizione di specie

1. INTRODUCTION

There are several well known Machine Learning techniques highly suitable for habitat modeling (Goethals & De Pauw 2001; Dakou *et al*. 2007). We applied classification trees and artificial neutral networks (ANNs) to model presence/absence of *Telestes muticellus*, in order to evaluate their reliability and to compare their performances.

2. STUDY AREA AND METHODS

The study system consisted of 198 sites located in Piedmont. *Teleste muticellus* was sampled in 139.

Twenty predictive environmental variables were measured. From these, thank to attribute-selection - stepwise procedure and Goldberg's (1989) genetic algorithm - we obtained a subset of 10 inputs. The presence/absence and the inputs data were obtained from the "*Monitoraggio della fauna ittica in Piemonte*" (Piedmont Region, 2006).

We used the ''Top-Down Induction of Decision Trees'' (Quinlan 1986) - J48 algorithm with binary split (Dakou *et al*. 2007). The tree post-pruning optimization (confidence factor 0.15 - 0.25) was applied. Ten fold cross-validation (Kohavi 1995) was repeated 10 times. We compared the performance of the unpruned and pruned trees with Mann-Whitney tests.

Three-layered feed-forward neural networks with bias, trained using the error backpropagation algorithm (Rumelhart *et al*. 1986), were built. The number of neurons was determined by trial and error (Bishop 1995). Momentum was set to 0.2, learning rate to 0.1. Nine fold cross-validation (Goethals *et al*. 2007) repeated 10 times was used. Mann-Whitney tests were performed to compare the performance of the ANNs and decision trees. Performance was assessed on the basis of: correctly classified instances, sensitivity, specificity, Cohen's k coefficient, and the area under the ROC curve.

3. RESULTS

The performance of the decision tree models was satisfactory; the optimal confidence level of pruning is 0.21. No significant differences in the predictive performance between pruned and unpruned trees were detected (Mann-Whitney tests). The final tree chosen had very good performances (CCI= 78.8%, sensitivity= 84.2%, specificity= 67.0%, Cohen's k= 0.47, area under the ROC curve= 0.7). The ANN architecture was: 10 input nodes (resulting from attribute selection), 6 hidden nodes, and 1 output node. ANNs have better performance than decision trees (Mann-Whitney tests: $p < 0.05$), except that for Cohen's k coefficient. The final ANN chosen had very good performances (CCI= 79.3%, sensitivity= 80.9%, specificity= 80.2%, Cohen's k= 0.43, area under the ROC curve= 0.8).

4. DISCUSSION AND CONCLUSIONS

Our results highlighted the need for different approaches and for the variable selection, which reduces the number of inputs, improves the predictive performance (D'heygere *et al*., 2003, 2006), and allows to obtain reliable models according to Cohen's k. The best performance of ANNs stressed their usefulness, as evidenced by their predictive power - not based on chance (Gabriels *et al*. 2007) - and good discrimination (Hosmer & Lemeshow 2000). Therefore, we strongly recommend to use different techniques, in order to select the best suited to manage each specific problem.

REFERENCES

Bishop C.M., 1995 - *Neural networks for pattern recognition*. *Oxford University Press*, Oxford, 504 pp.

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

D'heygere T., Goethals P.L.M. & De Pauw N., 2003 - Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. *Ecol*. *Mod*., 160: 291-300.

D'heygere T., Goethals P.L.M. & De Pauw N., 2006 - Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. *Ecol*. *Mod*., 195: 20-29.

Gabriels W., Goethals P.L.M., Dedecker A.P., Lek S. & De Pauw N., 2007 - Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks. *Aquatic Ecol*., 41: 427-441.

Goethals P. & De Pauw N., 2001 - Development of a concept for integrated ecological assessment in Flanders, Belgium. *J*. *Limnol*., 60: 7-16.

Goethals P.L.M., Dedecker A.P., Gabriels W., Lek S. & De Pauw N., 2007 - Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquatic Ecol*., 41: 491-508.

Goldberg D.E., 1989 - *Genetic algorithm in search, optimization and machine learning*. Addison-Winsley Publishing Company, Reading, 412 pp.

Hosmer D. & Lemeshow S., 2000 - *Applied Logistic Regression*,. John Wiley & Sons Inc., New York, 392 pp.

Kohavi R., 1995 - A study of cross-validation and bootstrap for estimation and model selection. In: Mellish C.S. (a cura di), *Proccedings of the 14th international joint conference on artificial intelligence*. Morgan Kaufmann Publisher, Montreal: 1137-1143.

Quinlan JR. 1986 - Induction of decision trees. *Mach*. *Learn*., 1: 81-106.

Rumelhart D.E., Hinton G.E. & Williams R.J., 1986 - Learning representations by back-propagation errors. *Nature*, 323: 533-536.