Environ. Eng. Res. Vol. 14, No. 2, pp. 102~110, 2009
Korean Society of Environmental Engineers

A Survey of Applications of Artificial Intelligence Algorithms in Eco-environmental Modelling

Kangsuk Kim and Joonhong Park†

School of Civil and Environmental Engineering, Yonsei University, Seoul 120-74, Republic of Korea

Received January 2009, accepted April 2009

Abstract

Application of artificial intelligence (AI) approaches in eco-environmental modeling has gradually increased for the last decade. Comprehensive understanding and evaluation on the applicability of this approach to eco-environmental modeling are needed. In this study, we reviewed the previous studies that used AI-techniques in eco-environmental modeling. Decision Tree (DT) and Artificial Neural Network (ANN) were found to be major AI algorithms preferred by researchers in ecological and environmental modeling areas. When the effect of the size of training data on model prediction accuracy was explored using the data from the previous studies, the prediction accuracy and the size of training data showed nonlinear correlation, which was best-described by hyperbolic saturation function among the tested nonlinear functions including power and logarithmic functions. The hyperbolic saturation equations were proposed to be used as a guideline for optimizing the size of training data set, which is critically important in designing the field experiments required for training AI-based eco-environmental modeling.

Keywords: Eco-environmental modeling, Data mining, Artificial intelligence, Decision Tree (DT), Artificial Neural Network (ANN), Training data, Prediction accuracy

1. Introduction

As disturbances and damages on eco-environmental systems by human activities become severe and widespread, conservation and restoration of the vital systems are growing concerns in sustainable development as well as environmental policy. This seems to be a global trend in these days. In making decision on sustainable development planning, basic eco-environmental information is required. Such basic eco-environmental information includes the diversity, abundance and distribution of biota as well as environmental quality.[1] Particularly, to examine whether a construction planning is eco-environmentally sound, such eco-environmental information is needed to be linked with geographic information as a form of maps. Because of these reasons, the needs for the acquisition and appropriate application of eco-environmental information are being increased.

The environment is a complex and dynamic system so that we have no simple sets of rules for describing that system at this time point. Also, it is impractical and inefficient approach that a lot of studies on eco-environmental problems and issues depend only on field measurement or experimentation.[2] Moreover, it is time-consuming and expensive work. Researchers have a variety of tools for collecting and analyzing data, but relatively few tools that facilitate eco-environmental reasoning and prediction.[3] For these reasons, mathematical models and computer simulations began to be used as the appropriate means to get more insight.[2] However, modelling of the eco- environmental systems using deterministic approach is often limited because such approach requires huge amounts of data for modeling ecological and environmental systems with natures of high complexity and nonlinearity. It may be more reasonable to use empirical approach to modeling of eco-environmental systems.

The fast-growing tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful data analysis tools. That has described as a 'data rich but information poor' situation.[4] Consequently, important decisions are often made based not on the information-rich data stored in databases but rather on a decision maker's intuition. The major reason that data mining has attracted a great deal of attention in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.[4] With the development of computer and information technology, data mining has became more

† Corresponding author
E-mail: parkj@yonsei.ac.kr
Tel: +82-2-2123-5798, Fax: +82-2-312-5798

popular due to its strong ability to predict unknown information using a training data set of previously-known information from a system of interest.[5,6] Data mining is a process of querying and extracting useful information, patterns, and trends often previously unknown from large quantities of existing data.[7] In data mining approach, particularly, artificial intelligence (AI) techniques (e.g., decision tree, artificial neural network, genetic algorithm, support vector machine, case-based reasoning and so far) facilitate ecological and environmental reasoning. The most immediate impact of AI technologies will be on the way of researchers to organize, develop, and implement models.[3]

Although the AI-based data mining methods were developed in the fields of statistics, computer science, and engineering, the experts of business administration, economics and information technology seem to be the major groups to apply these methods in aids in their decision making processes.[8] In these days, AI algorithms and their applications are considered as well-established tools in medical, pharmaceutical, and biological research areas as well. However, only a limited number of AI-applications were reported in eco-environmental field at the early 1990s.[9-13] In this study, we attempted to survey the current uses of AI in ecological and environmental modeling, with special emphases on examining in which AI algorithms were mainly used in various environmental and ecological research areas. In addition, to propose a guideline for designing the size of training data set for ecological and environmental AI-modeling, prediction accuracy in response to size of training data set was investigated using the available data from literature. Nonlinear correlation equations were proposed to describe the relationship between model accuracy and the size of training data set. In this work, the statistical analysis was conducted only with supervised algorithms since measured target values are needed in training ANN and DT algorithms.

2. AI-technologies in Data Mining Approach

2.1. Basic Principles

Data mining has been defined as 'the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical technique.[14] Data mining involves an integration of techniques from multiple disciplines such as statistics, database technology, pattern recognition, machine learning, and other areas[15] and also has contribution from many other technologies. One such technology is machine learning (algorithms that improve their performance automatically through experience). Machine learning has roots in artificial intelligence, popularly known as AI.[7]

AI is a branch of computer science that is principally concerned with using computational models to understand how human think and behave.[16] AI-technologies have played a major role in data mining and may provide the high speed, computational tools and techniques.[3] Various AI techniques are used for association, estimation, classification, prediction and segmentation, yet each AI technique has its distinct strength and

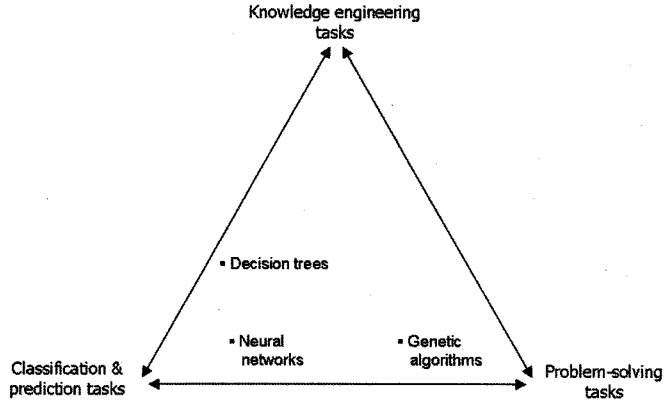

Fig. 1. AI-techniques in a simplex of three major data mining tasks.[18]

high performance in specific fields. For instance, Moustakis et al.[17] identified three major tasks (factors): (i) knowledge engineering task - acquisition of expert knowledge and its refinement to gain additional knowledge (e.g. mining of such deductive databases by inductive logic programming); (ii) problem solving (e.g. scheduling, optimization, etc.); (iii) classification and prediction, the association of these techniques when viewed in terms of the simplex these factors, as remapped by Adriaans & Zantinge,[18] displayed in Fig. 1. More techniques could be added in Fig. 1. Several powerful and popular AI-based data mining techniques, such as decision tree, artificial neural network and so far, are described in following sub sections.

2.2. Decision Tree

Decision tree (DT) is a powerful and popular tool for classification and prediction. DT is a non-parametric modelling approach, which consists of recursive partitions of the multidimensional space defined by the predictors into groups that are as homogenous as possible in term of the response.[11,19] The result of the analysis is a binary hierarchy structure called a decision tree with branches and leaves that contains the rules to predict the new cases.[6,19] (Fig. 2)

DT has many advantages over other model approaches.[7,11] Namely, (1) it has no strict assumption for the distribution of the target variable. (2) It deals with non-linear models easily without any variable transformation. (3) It also typically requires less training time compared to other AI techniques, such as artificial neural networks and support vector machines, while attaining similar accuracies.[20] (4) It can clearly indicate the relative importance of input variables. (5) Finally, the analyst can easily interpret a DT because it can generate understandable rules. It is not a 'black box' like the neural networks. Naturally, DT also has its limitations. (1) It requires a relatively large amount of training data. (2) It cannot express linear relationships in a simple and concise way. (3) It cannot produce a continuous output due to its binary nature. (4) It has no unique solution, that is, there is no best solution.[10,12]

For DT analysis, various algorithms, such as CHAID[21], CART,[19] and C4.5[22] have been proposed. In recent, improved algorithms with combining their merits are introduced and commercialized by researchers and software.

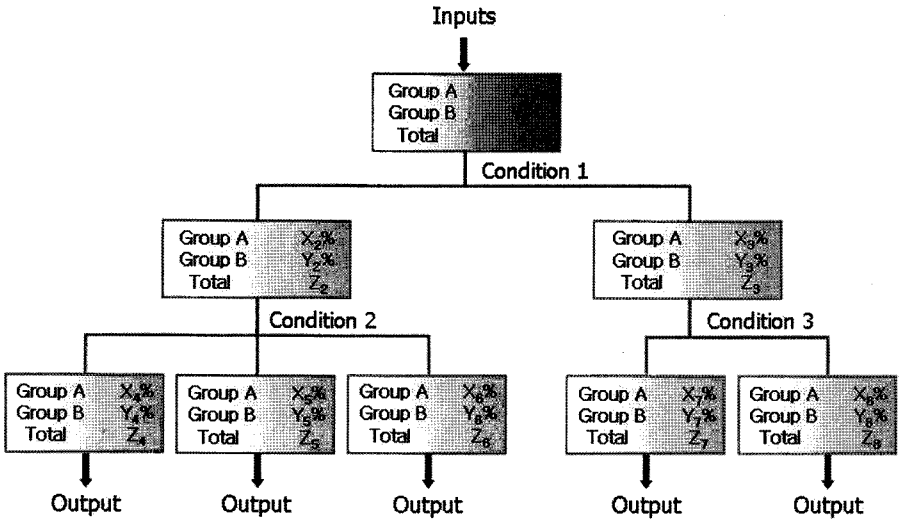

Fig. 2. General structure of decision tree (DT).

2.3. *Artificial Neural Network*

Artificial neural network (ANN) is an AI-technique that attempts to mimic the human brain's problem solving capabilities.[23] ANN model share the human brains capacity to learn from repeated number of inputs, by adjusting the weights that are assigned to the neurons (Fig. 3). ANNs are capable of self-organization and learning; patterns and concepts can be extracted directly from historical data.[24]

ANNs have recently become the focus of much attention, largely because of their wide range of applicability and the case with which they can treat complicated problems. These make powerful tools for models, especially when the underlying data relationships are previously unknown or complex nonlinear even if the data are imprecise and noisy.[9] In general, ANNs can be applied to the following types of problems: pattern classification, clustering and categorization, function approximation, prediction and forecasting, optimization, associative memory, and process control.[25]

ANN technique holds many advantages over conventional modelling methods. With respect to data processing, the type of relationship between the input and output data is determined purely from the information presented, with no presumptions from the network.[26] In addition, it is fault-tolerant both in model development and in subsequent applications; discontinuities in the data, different levels of data precision, noise, and data scatter are easily accommodated.[27] It is also extremely fast and flexible; advances in computing power have minimized the time required to develop models, as well as the time required to retrain models to incorporate new data and to reflect process modifications.[24]

With respect to the disadvantages of the ANN modelling technique, many researchers consider the developed models to be "black-box" models, as ANNs do not yield explicit rules.[14,26] This is the biggest criticism directed at ANNs. In addition, little is known about the applicability of the models to data that lie outside the domain on which the models were trained. No set

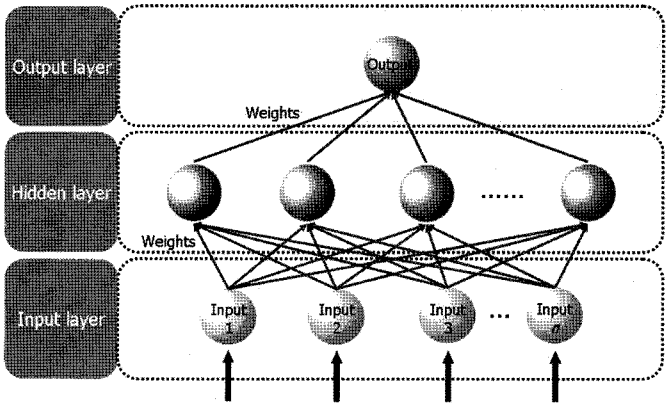

Fig. 3. General structure of artificial neural network (ANN).

protocol for developing ANN models exists; each modeler may incorporate different modelling techniques. Finally, it is data intensive and is best suited to problems where large data sets exist.[28] Thus ANN is good choice for most classification and prediction tasks when results of the model are more important than understanding how the model works.[14]

A variety of ANN algorithms have been proposed. At present, two popular ANN algorithms are (1) multi-layer feed-forward neural networks trained by backpropagation algorithm, i.e. backpropagation network (BPN), and (2) Kohonen self-organizing mapping, i.e. Kohonen network (SOM). The BPN is most often used, but other algorithms have also gained popularity.[9] The choice of ANN algorithm depends on the nature of the problem to be solved.

2.4. *Other AI-techniques*

Besides decision tree and artificial neural network, AI techniques that recently have received considerable attention are genetic algorithm, support vector machine, case based reasoning.

Support vector machine (SVM) is a computer algorithm that can perform pattern recognition tasks and has its roots in statistical learning theory by Vapnik.[29] This technique has shown

promising empirical results in many practical applications, from handwritten digit recognition to text categorization.[8] SVM has also been successfully applied to an increasingly wide variety of biological applications. A common biomedical application of SVM is the automatic classification of microarray gene expression profiles.[30] In addition, SVM works very well with high-dimensional data and avoids the curse of dimensionality problem. Another unique aspect of this approach is that it represents the decision boundary using a subset of the training samples, known as the support vectors.[8]

Genetic algorithm (GA) is a stochastic optimization technique first proposed and investigated by Holland[31] and a search algorithm based on the 'survival of the fittest' among string structure.[32] They applied the idea from biology research to guide the search to an optimal solution.[33] The general idea was to maintain an artificial ecosystem, consisting of a population of chromosomes. GA is particularly suitable for multi-parameter optimization problems with an objective function subject to numerous hard and soft constraints. It performs the search process in four stages: initializations, selection, crossover, and mutation.[33,34]

Case-based reasoning (CBR) is a problem solving technique that reuses past, similar cases to find solutions to problems.[35] It provides a solution to a new problem or situation case by referencing a case base (library of stored old cases). It mirrors the problem-solving approaches taken by human beings who solve current problems using past experiences. CBR just refers to specific knowledge of previously experienced situations while most artificial intelligence approaches depends on general knowledge of a problem domain. Thus, there is no possibility for overfitting.[36-38]

3. The Current Uses of AI in Eco-environmental Modelling

We aimed at providing an overview on the range of the current uses of AI in eco-environmental modeling. The previously reported studies applying AI-techniques to eco-environmental modelling were reviewed. As most of the previous studies reviewed were conducted between mid-1990s and present, just a decade has passed since use of AI algorithms began to be activated in eco-environmental modelling. We summarized the representative studies into 2 categories: ecological applications (Table 1) and environmental applications (Table 2).

As described in ecological applications (Table 1), we divided the subjects of ecological applications into plant, animal, and

Table 1. Summary of previously reported ecological studies using AI algorithms

Category	Application	Algorithm	Model accuracy*	No. training samples	No. input variables	Reference
Plant ecology	Distribution of vegetation on climate change	ANN	0.75	75,000	14	Hilbert et al.[39] (2001)
	Distribution and abundance of tree species following climate change	DT	0.46	1,700	33	Iverson et al.[10] (1998)
	Species distributions of vegetation	DT	0.59	410	25	Vayssieres et al.[11] (2000)
	Functional group abundance in a pasture ecosystem	DT	0.75	1,219	23	Zhang et al.[40] (2005)
	Tropical vegetation types and change detection in complex neotropical environments	DT	0.83	<700	12	Sesnie et al.[41] (2008)
Animal ecology	Production/biomass (P/B) ratio of Benthic invertebrate populations	ANN	0.80	750	13	Brey et al.[42] (1996)
	Benthic macroinvertebrate communities	ANN	N/R**	99	N/R	Chon et al.[43] (1996)
	Aquatic macroinvertebrate diversities	ANN	0.69	500	34	Park et al.[44] (2003)
	Trout abundance in rivers	ANN	0.88	N/R	8	Lek et al.[45] (1996)
	Riverine fish diversity	ANN	0.93	183	3	Guegan et al.[46] (1998)
	Abundance and diversity of hydrophilous Collembola in a riparian habitat	ANN	0.85	83	7	Lek-Ang et al.[47] (1999)
	Aquatic insect species richness	ANN	0.61	130	4	Park et al.[48] (2003)
	Algal blooms	ANN	N/R	N/R	7-11	Recknagel et al.[49] (1997)
	Primary production of phytoplankton in marine system	ANN	0.61	100	12	Scardi et al.[50] (1999)
	River phytoplankton dynamics	ANN	N/R	361	27	Jeong et al.[51] (2006)
Microbial ecology	Soil microbial diversity in a forest region	DT	0.61	137	7	Kim et al.[52] (2008)

*Model accuracy was expressed as either hit rate for DT or correlation coefficient (R^2) for ANN.
** N/R indicates a not reported value in the reference.

Table 2. Summary of previously reported environmental studies using AI algorithms

Category	Application	Algorithm	Model accuracy*	No. training samples	No. input variables	Reference
Greenhouse climate	Greenhouse climate control	ANN	0.74	N/R**	7	Seginer[53] (1997)
	Greenhouse climate control	ANN	0.97 0.97	509 808	5 6	Linker et al.[54] (1998)
Landscape	Land cover classification	DT	0.85	2,000	30	Pal et al.[20] (2003)
Water quality	Water quality parameters (salinity)	ANN	46, 47, 53 (Sensitivity)	<50,000	51, 69, 141	Maier et al.[55] (1996)
	Water quality management for river basin planning and water pollution control	ANN	Study on weights	70	3	Wen et al.[56] (1998)
	Nitrate leaching in agricultural drainage effluent	ANN	0.88	N/R	12	Kaluli et al.[57] (1998)
	Water quality	ANN	0.79	200	N/R	Schleiter et al.[58] (1999)
	Water quality and drinking water treatment process	ANN	0.79-0.95	17, 160-180	8-12	Baxter et al.[23] (2001)
	Stream channel stability	DT	ROC curve	N/R	15	Moret et al.[59] (2006)

*Model accuracy was expressed as either hit rate for DT or correlation coefficient (R^2) for ANN.
**N/R indicates a not reported value in the reference.

microbial ecology. The studies on plant ecology were mostly intended for modelling distribution and abundance of tree species. For the studies on animal ecology, the subjects of application were limited at aquatic invertebrate, fish, aquatic insect, and plankton. There were no studies targeting macro-size animals like mammals, amphibians, and reptiles. It may be the reason that acquisition of field-measured data for macro-size animals is not easy due to their highly mobile and dynamic nature. Regarding AI application in microbial ecology, a very limited number of studies were previously reported.[52,60,61] They proposed model framework for applying AI (DT models) for predicting soil microbial diversity in a Korean forest area. Because sampling of microorganisms might be relatively easy compared with animals' cases and the development of modern molecular tools facilitates rapid and quantitative analysis from field samples,[62,63] AI-based data mining approach may be highly applicable in microbial ecology studies.

In the current uses of AI in ecological modeling, DT was preferred for plant ecology while ANN was preferred for animal ecology. In the case of microbial ecology, it is difficult to make any generalization since the previously reported studies were from a single case (e.g., a Korean forest area). The rationale of choice of AI algorithm was not well described in the literature. The preference in AI algorithm probably resulted from manneristic choice of researchers rather than based upon algorithm's nature. Nevertheless, the studies for animal ecology seem to have a rationale for choosing ANN rather than DT. According to comparative analysis for training set size and number of input variables, ANN models were used with less number of training samples and input variables than those of DT models.[64,65] Probably, the previous researchers considered that ANN may be more efficient for modeling animal ecology, in which a greater size of training data set may be required for modeling its more mobile and dynamic features.

As described in environmental applications (Table 2), a wide range of research areas including greenhouse climate, landscape and water quality were covered by the current AI applications. Applications on water quality were dominant and its preferred algorithm was ANN in environmental studies. The model accuracies of environmental studies tend to be slightly high in comparison with those of ecological studies. Generally, abiotic environmental factors are easier to accurately and precisely measure in field conditions than ecological factors. This may explain the differences of the accuracies among ecological and environmental studies. Generalization of these findings should be carefully considered since this survey was carried out with a fairly limited number of previous studies. Nevertheless, these findings provided a rough guideline for selecting an algorithm in AI-based modeling of ecological and environmental phenomena.

4. Model Prediction Accuracy in Response to the Size of Training Data Set

Model accuracy is a permanent challenge to eco-environmental modelling.[66] The model accuracy is influenced by lots of factors such as kind of model and algorithm, data quality, training set size, data partitioning ratio, number of input variables, and so forth. Researchers have reported that training set characteristics, especially overall size in terms of the number of training samples, have a major effect on the model performance.[27,67,68] The results of investigations on relationship between model accuracy and training set size show that the model accuracy tends to improve as the training set size is enlarged. For instance, Foody et al.[65] assessed the effect of variations in the training set size on the classification accuracy of remotely sensed data sets by an artificial neural network. The results indicated that the accuracy increased significantly as a result of increasing the number of training cases. Pal et al.[20] using a decision tree model for land cover classification, also showed the accuracy of a decision tree model improves as the training set size is increased.

Besides of these, there are some other studies[27,69,70] which showed similar trends as mentioned above. However, very little is known about quantitative information on a minimal size of training data set to satisfy the target accuracy.

To gain such information, a good correlation between model accuracy and the size of training data set has yet to be sought for. For this purpose, a set of data for DT model accuracy in response to a size of training data set was obtained from Pal et al.,[20] and four equations (linear, hyperbolic saturation, logarithmic and power) were tested using regression analysis with the data points (Figure 4).

Linear: $Y = aX + b$ (1)

Hyperbolic saturation: $Y = cX / (d + X)$ (2)

Logarithmic: $Y = eln(X) + f$ (3)

Power: $Y = gX^h$ (4)

Where X is a number of training samples per input variable and Y is model prediction accuracy. a, b, c, d, e, f, g and h are coefficients of each function. Especially, coefficient c means asymptotic maximal accuracy (Y_{max}) and coefficient d means a number of training samples per input variable when $Y = 0.5Y_{max}$.

In results, all the tested equations except the linear equation were well-fitted. This indicates that nonlinear equations are good at describing the correlation between model accuracy and the size of training data set. Among the tested nonlinear equations, the hyperbolic saturation equation (R^2 = 0.965) was the most suitable for describing the effect of training set size on accuracy (Table 3). In the hyperbolic saturation equation, the value of coefficient c means that the estimated maximal accuracy is 0.866. Power and logarithmic equations showed underestimated values when model accuracy values are below 0.84. In this range of accuracy, the power and logarithmic model predictions showed lower values than hyperbolic saturation model predictions. When the values of accuracy are higher than 0.84, the hyperbolic saturation model predictions showed lower values than the power and logarithmic model predictions. These findings were supported by the results from the following nonlinear regression analysis (data not shown) in which the same nonlinear equations were tested with the literature data points from the eco-environmental studies using DT (Table 1 and 2).

To explore the effect of the size of training data set on ANN model accuracy, a set of data for ANN model accuracy in response to a size of training data set was obtained from previous eco-environmental modeling studies (Table 1 and 2), and four equations (linear, hyperbolic saturation, logarithmic and power) were also tested using regression analysis with the data points (Figure 5). Among the ANN studies listed in Table 1 and 2, some did not provide information on sampling size, external validation

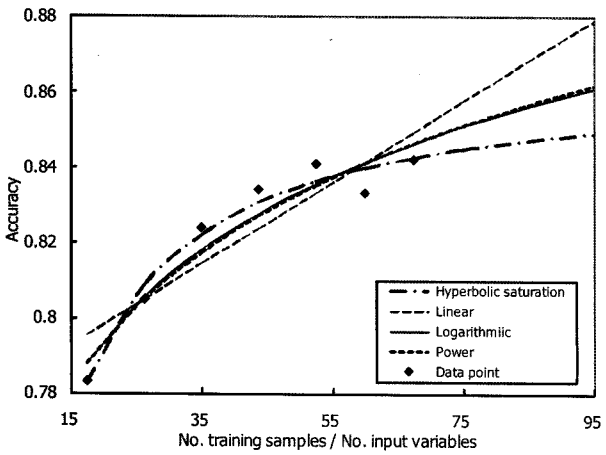

Fig. 4. Regression analysis of model accuracy and the size of training data for DT application. The data points were obtained from Pal et al.[20]

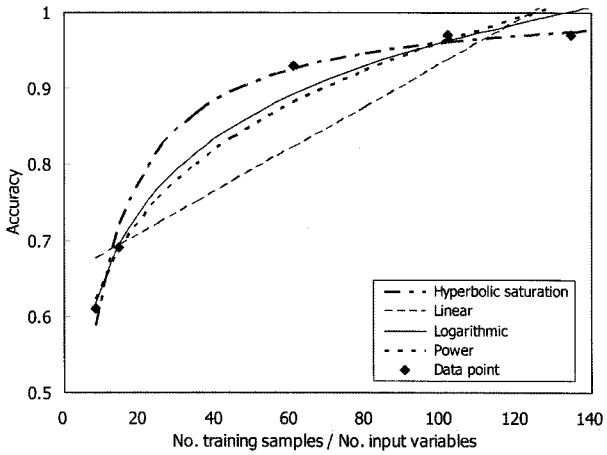

Fig. 5. Regression analysis of model accuracy and the size of training data for ANN applications. The data points were obtained from some of the previous studies of eco-environmental modeling.

Table 3. Summary of regression analysis for the effect of size of training data set on model accuracy

Regression type	Equation	DT		ANN	
		Coefficient (95% CI*)	R^2	*Coefficient* (95% CI)	R^2
Linear	$Y = aX + b$	a = 0.0011 ± 0.0006 b = 0.7765 ± 0.0278	0.810	a = 0.003 ± 0.002 b = 0.652 ± 0.193	0.827
Hyperbolic saturation	$Y = cX / (d + X)$	c = 0.8657 ± 0.0107 d = 1.858 ± 0.436	0.965	c = 1.019 ± 0.055 d = 6.115 ± 1.824	0.987
Logarithmic	$Y = eln(X) + f$	e = 0.0432 ± 0.0144 f = 0.6643 ± 0.0543	0.922	e = 0.138 ± 0.038 f = 0.324 ± 0.146	0.978
Power	$Y = gX^h$	g = 0.6782 ± 0.0461 h = 0.053 ± 0.018	0.918	g = 0.439 ± 0.119 h = 0.170 ± 0.064	0.964

*CI = confidence interval.

results, etc. These data were excluded when performing regression analysis with the different four equations. All the tested equations except the linear equation were well-fitted. This indicates that nonlinear equations are good at describing the correlation between ANN model accuracy and the size of training data set. Among the tested nonlinear equations, the hyperbolic saturation equation (R^2 = 0.987) was the most suitable for describing the effect of training set size on accuracy (Table 3).

The hyperbolic saturation equation was able to well capture the curve trend. Power and logarithmic equations showed underestimated values when model accuracy values are below 0.95. In this range of accuracy, the power and logarithmic model predictions showed lower values than hyperbolic saturation model predictions. When the values of accuracy are higher than 0.95, the hyperbolic saturation model predictions showed lower values than the power and logarithmic model predictions. These findings are similar to those observed in DT accuracy in response to the size of training data set.

Also, the results showed that all nonlinear model functions except the linear model were well-fitted and the hyperbolic saturation model (R^2 = 0.987) was the best for describing the effect of training set size on accuracy (Table 3). In a hyperbolic saturation function, the *c* and *d* coefficient values indicate (i) its maximal accuracy and (ii) the size of training data set for satisfying 50% of its corresponding *c* value, respectively. The *c* value for ANN (1.019 ± 0.055) is closer to a perfect accuracy (i.e., 1.0) than that for DT (0.8675 ± 0.0107). The *d* value for ANN (6.115 ± 1.824) is higher than that for DT (1.858 ± 0.436). These results suggest that ANN can achieve better accuracy when a larger size of training data set while DT can achieve better accuracy when a smaller size of training data set. Model simulation with the estimated c and d values was performed in the range of target accuracy between 0.50 and 0.95 because model accuracy below 0.50 dose not have practical meaning (Table 4). According to this simulation, ANN requires smaller sizes of samplings for satisfying the similar level of target accuracy.

Table 4. Model (hyperbolic saturation) simulations of the size of training data per input variable required for a wide range of model prediction accuracy in eco-environmental modelling using DT and ANN algorithms

Algorithm	Target* accuracy	Size of training data per input variable
DT	0.95	868.69
	0.90	158.64
	0.85	82.90
	0.80	53.94
	0.75	38.64
	0.70	29.18
	0.65	22.75
	0.60	18.10
	0.55	14.58
	0.50	11.82
ANN	0.95	84.19
	0.90	46.25
	0.85	30.76
	0.80	22.34
	0.75	17.05
	0.70	13.42
	0.65	10.77
	0.60	8.76
	0.55	7.17
	0.50	5.89

*Target accuracy was expressed as either hit rate for DT or correlation coefficient (R^2) for ANN.

5. Conclusion

The AI-based data mining obviously provides an attractive alternative approach for analyzing eco-environmental data and for modelling due to their specific features, such as non-linearity, adaptivity (i.e., learning from examples), and generalization. Also, it can reasonably simplify the complex eco-environmental systems with low measuring and computing effort but considerable accuracy. In this study, we reviewed the previous studies that used AI-techniques in eco-environmental modelling for examining the scope of such applications and which AI algorithms were mainly used. Representative studies were summarized into 2 categories: ecological applications and environmental applications. According to the results, DT and ANN were found to be major AI algorithms preferred by researchers in eco-environmental modelling areas. This preference in AI algorithms probably resulted from manneristic choice of researchers rather than based upon discriminated features of algorithms. This work improves our understanding of the current status and trend of AI-applications in eco-environmental modeling.

In addition, this review study allowed us to explore the statistical correlation between model prediction accuracy and the size of training data set. According to the statistical analysis, the prediction accuracy and the size of training data showed nonlinear correlation, and such correlations for DT and ANN were found to be well-described by the hyperbolic saturation equations. For training AI-based eco-environmental modeling, sampling from field works is required, and optimizing the size of field sampling is critically important. Because of this reason, the findings from this work will be used a guideline in design of an optimal size of field sampling for training AI-based eco-environmental modeling.

Acknowledgements

This subject is supported by Korea Ministry of Environment as "The Eco-technopia 21 project."

References

1. Geneletti, D., "Biodiversity impact assessment of roads: an approach based on ecosystem rarity," *Environmental Impact Assessment Review*, **23**(3), 343-365 (2003).
2. Cortes, U., "Artificial intelligence and environmental decision support systems," *Applied intelligence*, **13**(1), 77-91 (2000).

3. Rykiel, E. J., "Artificial intelligence and expert systems in ecology and natural resource management," *Ecological Modelling*, **46**(1-2), 3-8 (1989).
4. Han, J. and Kamber, M., Data mining: Concepts and techniques, Morgan Kaufmann Publishers, San Francisco (2001).
5. Witten, I. and Erank, E., Data mining: Practical machine learning tools and techniques with java implementations, Morgan Kaufmann Publishers, San Francisco (2000).
6. Dunham, M. H., Data mining: Introduction and advanced topics. Pearson Education Inc., Upper Saddle River (2002).
7. Thuraisingham, B., Data mining: Technologies, techniques, tools, and trends. CRC press, New York (1999).
8. Tan, P.-N., Steinbach, M., and Kumar, V., "Introduction to data mining," Addison-Wesley, Boston (2005).
9. Lek, S. and Guegan, J. F., "Artificial neural networks as a tool in ecological modelling, an introduction," *Ecological Modelling*, **120**(2-3), 65-73 (1999).
10. Iverson, L. R. and Prasad, A. M., "Predicting abundance of 80 tree species following climate change in the eastern United States," *Ecological Monographs*, **68**(4), 465-485 (1998).
11. Vayssieres, M., "Classification trees: An alternative non-parametric approach for predicting species distributions," *J. Vegetation Science*, **11**(5), 679-694 (2000).
12. Scheffer, J., "Data mining in the survey setting: Why do children go off the rails?," *Res. Letters in the Information and Mathematical Sciences*, **3**, 161-189 (2002).
13. Yang, C.-C., Prasher, S. O., Enright, P., Madramootoo, C., Burgess, M., Goel, P. K., and Callum, I., "Application of decision tree technology for image classification using remote sensing data," *Agricultural Systems*, **76**(3), 1101-1117 (2003).
14. Berry, M. J. A. and Linoff, G., Data mining techniques for marketing, sales, and customer support, John Wiley & Sons, New York (1997).
15. Hand, D. J., "Data mining: Statistics and more?," *The American Statistician*, **52**(2), 112-118 (1998).
16. Tanimoto, S. L., The elements of artificial intelligence, Computer Science Press, Rockvile (1987).
17. Moustakis, V. S., Lehto, M., and Sale\vendy, G., "Survey of expert opinion: which machine learning method may be useful for which task?," *Internat. J. Human-Computer Interaction*, **8**(3), 221-236 (1996).
18. Adriaans, P. W. and Zantinge, D., Data mining, Addison Wesley Longman, Harlow (1996).
19. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, J. C., Classification and regression trees, The Wasworth Statistics/Probability Series, Chapman & Hall, New York (1984).
20. Pal, M. and Mather, P. M., "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote Sensing of Environment*, **86**(4), 554-565 (2003).
21. Kass, G. V., "An exploratory technique for investigating large quantities of categorical data," *Appl. Statistics*, **29**(2), 119-127 (1980).
22. Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo (1993).
23. Baxter, C. W., Zhang, Q., Stanley, S. J., Shariff, R., Tupas, R. R., and Stark, H. L., "Drinking water quality and treatment: The use of artificial neural networks," *Canadian Journal of Civil Engineering*, **28**(1), 26-35 (2001).
24. Baxter, C. W., Stanley, S. J. and Zhang, Q., "Development of a full-scale artificial neural network model for the removal of natural organic matter by enhanced coagulation," *Aqua*, **48**(4), 129-136 (1999).
25. Jain, A., "Artificial neural networks: A tutorial," *Computer*, **29**(3), 31 (1996).
26. Harvey, S., "An introduction to artificial intelligence," *Appita Journal*, **51**(1), 20-24 (1998).
27. Foody, G. M. and Arora, M. K., "An evaluation of some factors affecting the accuracy of classification by an artificial neural network," *International Journal of Remote Sensing*, **18**(4), 799-810 (1997).
28. Zhang, Q. and Stanley, S. J., "Forecasting raw-water quality parameters for the North Saskatchewan River by neural network modeling," *Water Research*, **31**(9), 2340-2350 (1997).
29. Vapnik V., Statistical learning theory, John Wiley & Sons, New York (1998).
30. Noble, W. S., "What is a support vector machine?," *Nature Biotechnology*, **24**(12), 1565-1567 (2006).
31. Holland, J. H., Adaptation in natural and artificial systems, University of Michigan Press, Ann Arbor (1975).
32. Goldberg, D. E., Genetic algorithms in search, optimization and machine learning, Addison-Wesley Publishing Co. Inc., MA, pp. 412 (1989).
33. Wong, F. and Tan, C., Hybrid neural, genetic, and fuzzy systems, In G. J. Deboeck (Ed), Trading on the Edge, John Wiley, New York, pp. 243-261 (1994).
34. Davis, L., Handbook of genetic algorithms, Van Nostrand Reinhold Publishers, New York (1991).
35. Kolodner, J. L., Case-based reasoning, Morgan Kaufmann Publisher, San Francisco (1993).
36. Watson, I., Applying case-based reasoning: Techniques for enterprise systems, Morgan Kaufmann Publishers, San Francisco (1997).
37. Shin, K. S. and Han, I., "Case-based reasoning supported by genetic algorithms for corporate bond rating," *Expert Systems with Applications*, **16**(2), 85-95 (1999).
38. Humphreys, P., McIvor, R., and Chan, F., "Using case-based reasoning to evaluate supplier environmental management performance," *Expert Systems with Applications*, **25**(2) 141-153 (2003).
39. Hilbert, D. W. and Ostendorf, B., "The utility of artificial neural networks for modelling the distribution of vegetation in past, present and future climates," *Ecological Modelling*, **146**(1-3), 311-327 (2001).
40. Zhang, B., Valentine, I., and Kemp, P. D., "A decision tree approach modelling functional group abundance in a pasture ecosystem," *Agriculture, Ecosystems & Environment*, **110**(3-4), 279-288 (2005).
41. Sesnie, S. E., Gessler, P. E., Finegan, B., and Thessler, S., "Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments,"

Remote Sensing of Environment, **112**(5), 2145-2159 (2008).
42. Brey, T., "Artificial neural network versus multiple linear regression: Predicting P/B ratios from empirical data," *Mar. Ecol. : Prog. Ser.*, **140**(1-3), 251-256 (1996).
43. Chon, T.-S., Park, Y. S., Moon, K. H., and Cha, E. Y., "Patternizing communities by using an artificial neural network," *Ecological Modelling*, **90**(1), 69-78 (1996).
44. Park, Y.-S., Verdonschot, P. F. M., Chon, T.-S., and Lek, S., "Patterning and predicting aquatic macroinvertebrate diversities using artificial neural network," *Water Res.*, **37**(8), 1749-1758 (2003).
45. Lek, S., Belaud, A., Baran, P., Dimopoulos, I., and Delacoste, M., "Role of some environmental variables in trout abundance models using neural networks," *Aquatic Living Resources*, **9**, 23-29 (1996).
46. Guegan, J.-F., Lek, S., and Oberdorff, T., "Energy availability and habitat heterogeneity predict global riverine fish diversity," *Nature*, **391**(6665), 382-384 (1998).
47. Lek-Ang, S., Deharveng, L., and Lek, S., "Predictive models of collembolan diversity and abundance in a riparian habitat," *Ecological Modelling*, **120**(2-3), 247-260 (1999).
48. Park, Y.-S., Cereghino, R., Compin, A., and Lek, S., "Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters," *Ecological Modelling*, **160**(3), 265-280 (2003).
49. Recknagel, F., French, M., Harkonen, P., and Yabunaka, K.-I., "Artificial neural network approach for modelling and prediction of algal blooms," *Ecological Modelling*, **96**(1-3), 11-28 (1997).
50. Scardi, M., and HardingJr, L. W., "Developing an empirical model of phytoplankton primary production: a neural network case study," *Ecological Modelling*, **120**(2-3), 213-223 (1999).
51. Jeong, K.-S., Kim, D.-K., and Joo, G.-J., "River phytoplankton prediction model by Artificial Neural Network: Model performance and selection of input variables to predict time-series phytoplankton proliferations in a regulated river system," *Ecological Informatics*, **1**(3), 235-245 (2006).
52. Kim, K., Ki, D., Son, I., Oh, K., and Park, J., " Application of artificial intelligence to planning sustainable construction: an ecological quality assessment of soil," in *Proceedings of the 7th International Conference on Sustainable Energy Technologies*, Korea Institute of Ecological Architecture and Environment, Seoul, pp. 1205-1212 (2008).
53. Seginer, I., "Some artificial neural network applications to greenhouse environmental control," *Computers and Electronics in Agriculture*, **18**(2-3), 167-186 (1997).
54. Linker, R., Seginer, I., and Gutman, P. O., "Optimal CO_2 control in a greenhouse modeled with neural networks," *Computers and Electronics in Agriculture*, **19**(3), 289-310 (1998).
55. Maier, H., "The use of artificial neural networks for the prediction of water quality parameters," *Water resources research*, **32**(4), 1013-1022 (1996).
56. Wen, C., "A neural network approach to multiobjective optimization for water quality management in a river basin," *Water Resour. Res.*, **34**(3), 427-436 (1998).
57. Kaluli, J., "Modeling nitrate leaching using neural networks," *Water Sci. Technol.*, **38**(7), 127-134 (1998).
58. Schleiter, I. M., Borchardt, D., Wagner, R., Dapper, T., Schmidt, K.-D., Schmidt, H.-H., and Werner, H., "Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks," *Ecological Modelling*, **120**(2-3), 271-286 (1999).
59. Moret, S. L., Langford, W. T., and Margineantu, D. D., "Learning to predict channel stability using biogeomorphic features," *Ecological Modelling*, **191**(1), 47-57 (2006).
60. Ki, D., Park, J., Lee, J., and Rho, P., "A weak correlation of field-determined soil microbial diversity with quantitative ecological map information and its methodological implication in estimation soil ecological quality," *KSCE*, **27**(6B), 703-710 (2007).
61. Ki, D., Kang, H. G., Lee, S. E., Heo, J., and Park, J., "Sensitivity analysis of the effect of soil ecological quality information in selecting eco-friendly road route," *Journal of Korean Society of Soil and Groundwater Environment*, **13**(3), 37-44 (2008).
62. Dunbar, J., Ticknor, L., and Kuske, C., "Phylogenetic specificity and reproducibility and new method for analysis of terminal restriction fragment profiles of 16S rRNA genes from bacterial communities," *App. Environ. Microbiology*, **67**(1), 190-197 (2001).
63. Tiedje, J. M., Asuming-Brempong, S., Nusslein, K., Marsh, T., and Flynn, S., "Opening the black box of soil microbial diversity," *App. Soil Ecology*, **13**(2), 109-122 (1999).
64. Hepner, G. F., Logan, T., Ritter, N., and Bryant, N., "Artificial neural network classification a minimal training set: Comparison to conventional supervised classification," *Photog. Eng. Remote Sensing*, **56**(4), 469-473 (1990).
65. Foody, G. M., Mcculloch M. B., and Yates, W. B., "The effect of training set size and composition on artificial neural network classification," *International J. Remote Sensing*, **16**(9), 1707-1723 (1995).
66. Recknagel, F., "Applications of machine learning to ecological modelling," *Ecological Modelling*, **146**(1-3), 303-310 (2001).
67. Foody, G. M., Mathur, A., Sanchez-Hernandez, C., and Boyd, D.S., "Training set size requirements for the classification of a specific class," *International J. Remote Sensing*, **104**(1), 1-14 (2006).
68. Kavzoglu, T., An investigation of the design and use of feed-forward artificial neural networks in the classification of remotely sensed images, *PhD thesis*, University of Nottingham, Nottingham (2001).
69. Leshno, M., and Spector, Y., "The effect of training data set size and the complexity of the separation function on neural network classification capability: The two-group case," *Naval Res. Logistics*, **44**(8), 699-717 (1997).
70. Landgrebe, D., "On the relationship between class definition precision and classification accuracy in hyperspectral analysis," in *Proceedings of IEEE Geoscience and Remote Sensing Symposium*, IEEE, Honolulu (2000).