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 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.

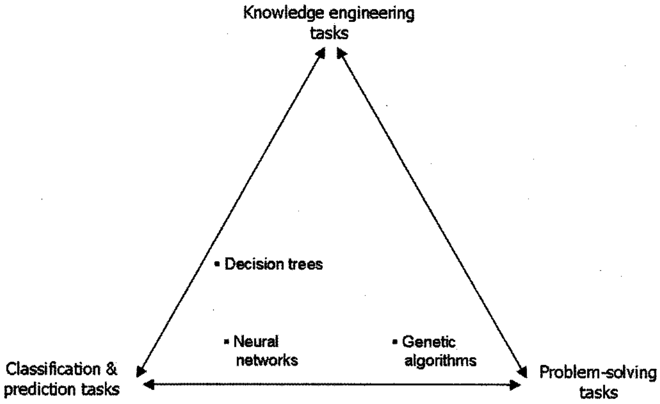

Fig. 1. AI-techniques in a simplex of three major data mining tasks.[18]

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.