A Short Review on the Application of Computational Intelligence and Machine Learning in the Bioenvironmental Sciences

Shinji Fukuda
Institute of Tropical Agriculture
Kyushu University
Fukuoka, Japan
shinji-fkd@agr.kyushu-u.ac.jp

Bernard De Baets
Faculty of Bioscience Engineering
Ghent University
Ghent, Belgium
bernard.debaets@ugent.be

Abstract—**This paper aims to provide a short review on the application of computational intelligence (CI) and machine learning (ML) in the bioenvironmental sciences. To clearly illustrate the current status, we limit our focus to some key approaches, namely fuzzy systems (FSs), artificial neural networks (ANNs) and genetic algorithms (GAs) as well as some ML methods. The trends in the application studies are categorized based on the targets of the model such as animal, fish, plant, soil and water. We give an overview of specific topics in the bioenvironmental sciences on the basis of the review papers on model comparisons in the field. The summary of the modelling approaches with respect to their aim and potential application fields can promote the use of CI and ML in the bioenvironmental sciences.**

Keywords-predictive modelling; knowledge-based modelling; spatiotemporal modelling; biology; environment

I. INTRODUCTION

Following the continuous development of computational intelligence (CI) and machine learning (ML), these modelling approaches are increasingly recognized and widely applied as a useful tool to analyze observation data obtained from biological and environmental surveys and experiments. CI is a group of modelling techniques whose primary aim is to realize a computer-based intelligent system on the basis of fuzzy systems (FSs), evolutionary algorithms (EAs) and artificial neural networks (ANNs). ML is a data-driven learner, aiming at achieving a higher prediction accuracy, and can be used for knowledge extraction. Indeed, a predictive model itself can be regarded as deduced knowledge (from data) to represent the input-output relationship of a target system. Finding a solution to the accuracy-interpretability trade-off is a research direction in both CI and ML communities, because of its usefulness for real-world applications. This paper aims to provide a brief overview of the application of CI and ML in order to facilitate the applications of advanced modelling techniques to solve problems in the bioenvironmental sciences in relation to biology, ecology and agriculture.

The structure of this study is as follows. In Section II, basic modelling approaches used in the bioenvironmental sciences, in general, are summarized. Then, the results on literature surveys on the applications of CI and ML are reported in Section III. After presenting some key review papers in the bioenvironmental science, we conclude this paper in Section IV.

II. BRIEF SUMMARY OF MODELLING APPROACHES IN BIOENVIRONMENTAL SCIENCE

In general, there are two modelling approaches to represent the responses of a target system to a given condition: process-based modelling and statistical correlative modelling. The view on these modelling approaches should be extended to the spatiotemporal dynamics of a target system. Consequently, we can categorize modelling approaches as well as real-world problems into three categories, namely predictive modelling, knowledge-based modelling, and spatiotemporal modelling (Fig. 1). The integration of these approaches can complement each other and have synergetic effects in practice. For instance, process-based models can be improved by the information obtained from data-driven machine learning methods (i.e., knowledge extraction and parameter identification), whereas prior knowledge on a system can improve predictions using machine learning methods. Here, we provide a brief summary of each modelling approach with a specific focus on the applications in the bioenvironmental sciences.

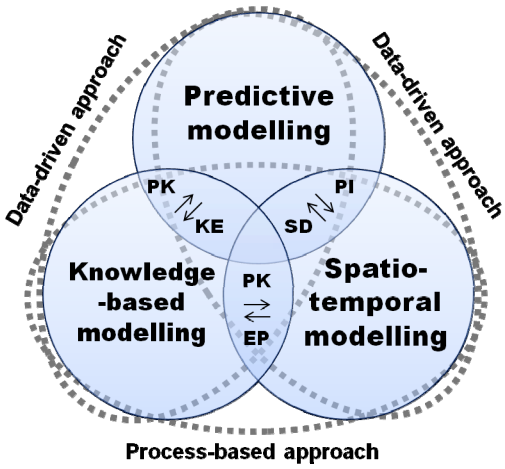

Figure 1. Conceptual framework of ecological modelling, in which PK stands for prior knowledge, KE is knowledge extraction, EP is emergent property, SD is system dynamics and PI is parameter identification.

This study was supported in part by the cooperation agreement between Ghent University and Kyushu University.

A. *Predictive Modelling*

To achieve a highly accurate model is an ideal and ultimate goal for modellers. The higher accuracy of various models has been achieved by means of various loss functions and learning or optimization algorithms. In addition to ML methods [1]–[2], ensemble methods have also been increasingly applied to ecological modelling [3]. In predictive modelling, the model accuracy should be tested using correctly designed cross-validation schemes. To cope with the accuracy-interpretability trade-off in CI and ML methods, several approaches have been proposed and applied to extract useful information from the model results (see [4] for ANNs and [5] for SVMs).

B. *Knowledge-based Modelling*

Knowledge-based models have been applied in order to incorporate expert knowledge or theoretical frameworks of a target system into its process-based models or simplified conceptual models. In the case of process-based models, the model accuracy depends largely on the availability of a detailed description and required data for the target system to be modelled. In contrast, conceptual models can be useful as tools to abstract the complex phenomena observed in real-world problems. Fuzzy rule-based models are one of the most applied methods for incorporating expert knowledge into models (see [6] for a review in agricultural engineering and [7] for a review in ecosystem management).

C. *Spatiotemporal Modelling*

In nature, most phenomena are dynamic in both time and space. It is, therefore, natural for a model to incorporate the spatiotemporal dynamics of a target system. For instance, environmental conditions such as temperature and rainfall show spatiotemporal dynamics, and then species distributions change according to their internal conditions and life stages. As such, considering spatiotemporal dynamics allows for analyses considering stochastic effects, which is typical for population dynamics modelling in ecology [8]. The accuracy or reliability of this approach depends largely on the model structure and numerical schemes to solve the governing equations of the model.

III. Computational Intelligence and Machine Learning in the Bioenvironmental Sciences

In this section, we report the results of a brief survey on the papers using CI and ML methods, namely ANNs, GAs, FSs, ML, CI, cellular automata (CA), classification and regression trees (CART), random forests (RF), support vector machines (SVMs) and evolutionary algorithms (EAs), which were registered in the ISI Web of Science® (see Appendix I). The literature survey was conducted using specific keywords as a target for each modelling approach, namely water, soil, animal, bird, fish, insect, bacteria and plant. Also, the number of publications per year was surveyed in order to grasp the trend. We then present an overview of applications of CI and ML methods based on some key review papers in agricultural engineering [6] and ecological modelling [8]–[10].

A. *Trends in the Published Papers*

Figs. 2 and 3 show the number of papers and proportion of methods used in papers, respectively, in which 5 keywords (CI, ML, FSs, GAs and ANNs) for methods and 8 keywords for the target of models (water, soil, animal, bird, fish, insect, bacteria

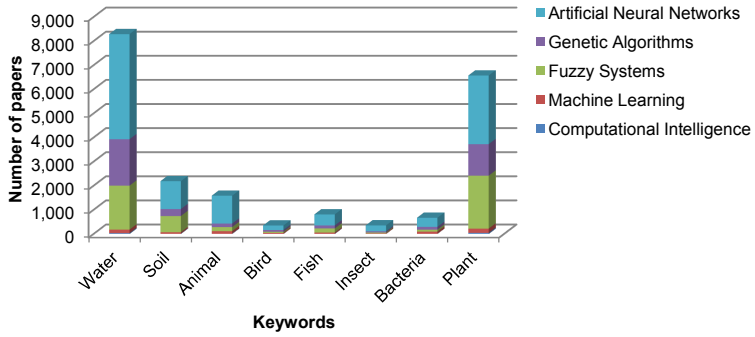

Figure 2. Number of papers regarding artificial neural networks, genetic algorithms, fuzzy systems, machine learning and computational intelligence in relation to specific keywords, namely water, soil, animal, bird, fish, insect, bacteria and plant.

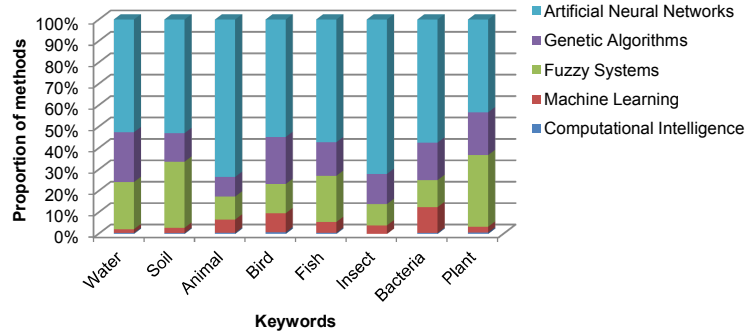

Figure 3. Proportion of methods in papers regarding artificial neural networks, genetic algorithms, fuzzy systems, machine learning and computational intelligence in relation to specific keywords, namely water, soil, animal, bird, fish, insect, bacteria and plant.

and plant) were used for the literature survey. In contrast, Figs. 4 and 5 show the number of papers and proportion of methods used in papers, respectively, in which 5 keywords (CA, CART, RF, SVMs and EAs) for methods and the same 8 keywords for the target of models were used. We can observe that the hottest topic for the application of CI and ML methods is "water," which is essential for all living beings. The need for a better use and management of water may have facilitated the application of advanced computational methods to solve complex and nonlinear problems. The keyword "plant" is the second largest number of papers, which may be partially because the word can also be used for industrial plants or power plants. Indeed, the application of CI and ML methods increased with their successes in industry (Fig. 6). The CI and ML methods are now increasingly applied in the bioenvironmental sciences, especially after the development of highly predictive ML methods such as SVMs (since 2000) and RF (in the last 5 years). The artificial intelligence methods (ANNs, GAs and FSs) have been actively employed since the early 1990s. CA has been gradually increasing with a stable number of applications since early 1990s. A similar trend can be observed for EAs. Although not considered here, the trend in hybrid model application should also be interesting and informative for the future application of CI and ML methods.

B. Key Review Papers in the Bioenvironmental Sciences

Agricultural engineering can be one of the fields with the highest potential for the application of CI and ML, because a

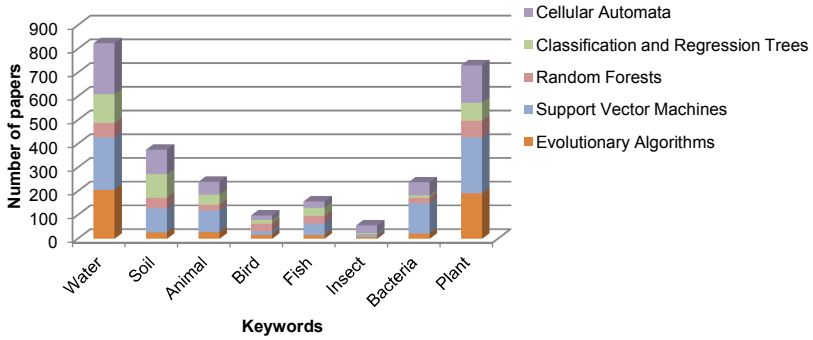

Figure 4. Number of papers regarding cellular automata, classification and regression trees, random forests, support vector machines and evolutionary algorithms in relation to specific keywords, namely water, soil, animal, bird, fish, insect, bacteria and plant.

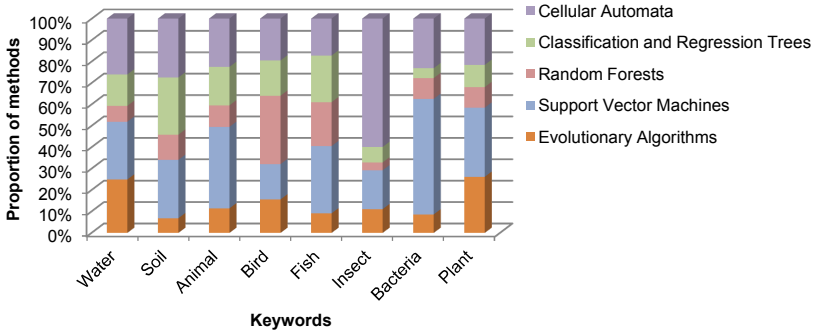

Figure 5. Proportion of methods in papers regarding cellular automata, classification and regression trees, random forests, support vector machines and evolutionary algorithms in relation to specific keywords, namely water, soil, animal, bird, fish, insect, bacteria and plant.

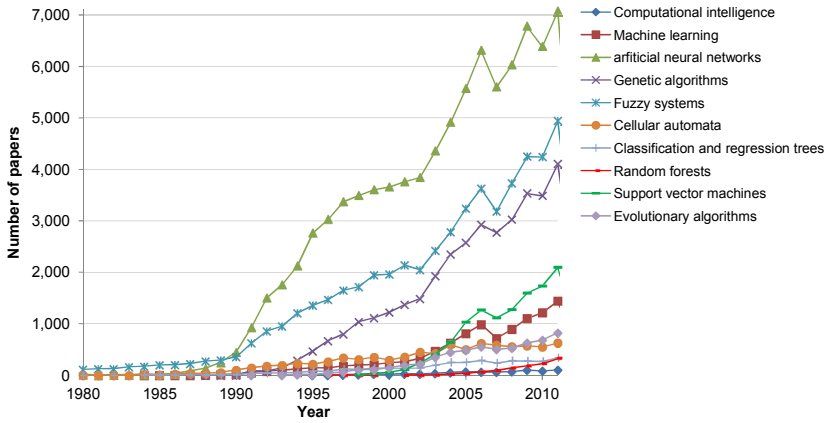

Figure 6. Number of papers per year published between 1980 and 2011 in the ISI Web of Science. The results shown are on the 10 keywords, namely computational intelligence, machine learning, artificial neural networks, genetic algorithms, fuzzy systems, cellular automata, classification and regressions trees, random forests, support vector machines and evolutionary algorimts.

lot of researchers use mathematical models, data mining tools and robots in order to improve management practices for their food production and distribution, and environment. Reference [6] reviewed the applications of fuzzy logic, ANNs, GAs, SVMs, Bayesian inference and Decision tree for each of crop management, irrigation, ET calculation, soil analysis, precision agriculture, and chemical application. SVMs were introduced as a new and promising method for classification and regression. In addition, hybrid models were also reviewed together with their applications.

Ecology was the first area of biology where quantitative models were constructed. Reference [8] is a review in which stochastic formulation of individual-based models (IBMs) and the deduction of population-level models (PLMs) from the IBMs is suggested. It provides an overview on the construction and analysis of stochastic models as well as three examples of ecological IBMs (i.e. metapopulation models, neutral models and spatial predator-prey models). In [8], the emergence of macroscopic dynamics was exemplified from three aspects: quasi cycle, spatial patterns and recurrent epidemics. From an application point of view, understanding possible responses of a system to a given condition (either mathematically or empirically) is useful for the interpretation of model results. In CI, multiagent models, artificial life and CA are similar research domains sharing the same goal for modelling. Despite the differences, the optimization of model parameters from observation data can be one of the interesting topics for both bioenvironmental, CI and ML scientists.

Ecoinformatics is a rapidly growing research field, which aims to generate new knowledge through innovative tools and approaches for discovering, managing, integrating, analyzing, visualizing, and preserving relevant biological, environmental, and socioeconomic data and information. In [9], the basic framework and workflow in ecoinformatics were reviewed, and then remaining challenges were discussed. As new ecological and environmental observational systems are moving ecology into the realm of big science, transdisciplinary collaboration will be needed to accumulate and analyze massive amounts of data in order to answer grand challenge questions in biological and environmental sciences. With regard to ecoinformatics, there are a series of review papers in the journal Ecological Informatics, which cover a wide range of topics: current trends and future challenges [11], information management [12], spatiotemporal ecological models [13], and hybrid modelling [14].

In parallel with [9], [10] overviewed integrated ecological-environmental management which requires a holistic view in practice. Reference [10] proposed to take seven steps, namely (1) define the problem, (2) determine the ecosystems involved, (3) identify causes and quantify all the sources to the problem, (4) set up a diagnosis to understand the relationship between the problem and the sources, (5) determine the tools we need to implement to solve the problem, (6) take the proper measures or actions and (7) follow the recovery process. Steps 1–3 may require expert knowledge in decision making. Steps 3–6 can be accomplished by means of predictive models and information retrieved from them. In steps 6–7, the decision making should be made considering the spatiotempral dynamics of target systems in response to the management actions under multiple criteria. In this regard, advanced CI and ML methods such as multiclass modelling and multiobjective optimization can be useful tools that can contribute to better planning and improved implementation of the management actions.

IV. Concluding Remarks

With an increasing amount of data and rapidly advancing CI and ML methods, research questions in the bioenvironmental sciences can be answered based on transdisciplinary research collaboration. The use of hybrid models or integration of several modelling approaches should be the key to solve problems that are, in general, complex and dynamic in space and time. The categorization of modelling approaches and target problems, according to predictive modelling, knowledge-based modelling, and spatiotemporal modelling, may help to identify a better solution to a specific problem when applying the CI and ML methods.

Appendix I

The following terms were used for the query in the ISI Web of Science® (Date of analysis: 28 June 2012).

- Artificial Neural Networks (ANNs): “neural networks” OR “neural network”
- Genetic Algorithms (GAs): “genetic algorithm” OR “genetic algorithms”
- Fuzzy Systems (FSs): “fuzzy”
- Machine Learning (ML): “machine learning”
- Computational Intelligence (CI): “computational intelligence”
- Cellular Automata (CA): “cellular automata” OR “cellular automaton”
- Classification And Regression Trees (CART): “classification and regression trees” OR “decision trees”
- Random Forests (RF): “random forest” OR “random forests”
- Support Vector Machines (SVMs): “support vector machine” OR “support vector machines”
- Evolutionary Algorithms (EAs): “evolutionary algorithm” OR “evolutionary algorithms” OR “evolutionary computation” OR “evolutionary optimization”

References

[1] C. Kampichler, R. Wieland, S. Calmé, H. Weissenberger, S. Arriaga-Weiss, “Classification in conservation biology: A comparison of five machine-learning methods,” Ecol. Inform., vol. 5, 441–450, 2010.

[2] R. Pino-Mejías, M.D. Cubiles-de-la-Vega, M. Anaya-Romero, A. Pascual-Acosta, A. Jordán-López, N. Bellinfante-Crocci, “Predicting the potential habitat of oaks with data mining models and the R system,” Environ. Model. Softw., vol. 25, 826–836, 2010.

[3] M.B. Araújo, M. New, “Ensemble forecasting of species distributions,” Trends Ecol. Evol., vol. 22, 42–47, 2007.

[4] J.D. Olden, M.K. Joy, R.G. Death, "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulation data," Ecol. Model., vol. 178, 389–397, 2004.

[5] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines," Mach. Learn., vol. 46, 389–422, 2002.

[6] Y. Huang, Y. Lan, S.J. Thomson, A. Fang, W.C. Hoffmann, R.E. Lacey, "Development of soft computing and applications in agricultural and biological engineering," Comput. Electron. Agric. vol. 71, 107–127, 2010.

[7] V. Adriaenssens, B. De Baets, P. Goethals, N. De Pauw, "Fuzzy rule-based models for decision support in ecosystem management," Sci. Total Environ., vol. 319, 1–12, 2004.

[8] A.J. Black, A.J. McKane, "Stochastic formulation of ecological models and their applications," Trends Ecol. Evol., vol. 27, 337–345, 2012.

[9] W.K. Michener, M.B. Jones, "Ecoinformatics: Supporting Ecology as a Data-Intensive Science," Trends Ecol. Evol., vol. 27, 85–93, 2012

[10] S.E. Jørgensen, S.N. Nielsen, "Tool boxes for an integrated ecological and environmental management," Ecol. Indic., vol. 21, 104–109, 2012.

[11] K.J. Metzger, R. Klaper, M.A. Thomas, "Implications of informatics approaches in ecological research," Ecol. Inform., vol. 6, 4–12, 2011.

[12] W.K. Michener, J. Porter, M. Servilla, K. Vanderbilt, "Long term ecological research and information management," vol. 6, 13–24, 2011.

[13] Q. Chen, R. Han, F. Ye, W. Li, "Spatio-temporal ecological models," Ecol. Inform., vol. 6, 37–43, 2011.

[14] L. Parrott, "Hybrid modelling of complex ecological systems for decision support: Recent successes and future perspectives," Ecol. Inform., vol. 6, 44–49, 2011.