Available online at www.elixirpublishers.com (Elixir International Journal)

Pollution

Elixir Pollution 42 (2012) 6420-6424

ISSN: 2229-712X

Application of artificial neural network in modelling of algal blooms an overview

Mageshkumar P[1] and Amal Raj S[2]

[1]Department of Civil Engineering, Kongu Engineering College, Erode 638052, India.

[2]Centre for Environmental Studies, Anna University Chennai, Chennai 600025, India.

ARTICLE INFO

Article history:

Received: 14 December 2011;

Received in revised form:

15 January 2012;

Accepted: 27 January 2012;

Keywords

Artificial Neural Network,

Algal Bloom,

Modelling Parameters.

ABSTRACT

Explosions like formation of algal blooms increasingly pollute both salt and fresh water ecosystems throughout the world. Owing to its negative impacts on human health and aquatic life, this widely reported phenomenon has become a serious environmental problem. While many process based, statistical and empirical models exist for water quality prediction, Artificial Neural Network (ANN) models are increasingly being used for water related applications because ANNs are often capable of modelling complex systems for which behavioral rules or underlying physical processes are either unknown or difficult to simulate. Theoretical knowledge about biological processes can be easily embedded into Neural Network models by means of a constrained training procedure. It is a complex non-linear function with many parameters that are adjusted (calibrated or trained) in such a way that the network output becomes similar to the measured output on a known data set. The feed forward neural network models are effective in predicting the non-linear behaviour of algalblooms and the correlation values are as high as 0.95 between the measured and calculated values. This paper presents an overview and provides a systematic approach for modelling the algal blooms using Artificial Neural Networks.

Introduction

Lakes have special importance owing to their value as natural ecosystems and centres of biodiversity, Role in sensing environmental pollution and value as a resource for water supply, hydropower, irrigation and amenity. Lakes are under increasing threat due to the nutrient enrichment from domestic and agricultural sources. Human have always been attached to lakes. Lakes are used by humans for many commercial purposes, including fishing, transportation, irrigation, industrial water supplies, and receiving waters for wastewater effluents. Lakes are the best available freshwater source on the Earth's surface. Aside from their importance for human use, lakes have intrinsic ecological and environmental values. With ecological modelling emerging as an invaluable tool for understanding the functioning of lakes, it is important to harness its potentials in ecosystem research. Though a number of models are available, the challenge is to identify the best possible model to be used for studying a lake ecosystem.

Algal Blooms

Aquatic ecosystems are very complex due to the diversity and connections of the components governing the system's dynamics. Explosion-like formations of algal blooms increasingly pollute both: salt and fresh water ecosystems throughout the world. Problems related to this nutrient over enrichment and excessive plant productions are probably the most common and have received public and scientific attention for the longest time. Owing to its negative impacts on human health and aquatic life, this widely reported phenomenon has become a serious environmental problem. This phenomenon results in numerous ecological and water quality changes in lakes. So it is essential to model the primary production in the lake or lagoon ecosystem.

An Overview of Artificial Neural Networks

Artificial Neural Networks (ANNs) have been widely used for modeling hydrological processes that are embedded with high non-linearity in both spatial and temporal scales. Further, ANNs are used as an effective approach for handling data in situations where the physical processes relationships are not fully understood and they are also well suited to modeling complex systems on a real-time basis.

The ANNs are functionally equivalent to a nonlinear regression model. ANNs can identify and learn correlated patterns between input data sets and corresponding target values. After training, ANNs can be used to predict the output of new independent input data. ANNs imitate the learning process of the animal brain and can process problems involving very nonlinear and complex data even if the data are imprecise. Thus they are ideally suited for the modelling of ecological data which are known to be very complex and often non-linear [6, 7, 14]. ANNs can benefit more from large samples than linear statistical models can. It is interesting to note that ANNs do not necessarily require a larger sample than is required by linear models in order to perform well.

Artificial neural networks (ANNs) are non-linear mapping structures based on the function of the human brain. They have been shown to be universal and highly flexible function approximators for any data. These make powerful tools for models, especially when the underlying data relationships are unknown.

Artificial neural networks (ANNs) are among the newest signal-processing technologies and the field is highly

Tele:

E-mail addresses: mageshee@gmail.com

interdisciplinary. An Artificial Neural Network is an adaptive, most often nonlinear system that learns to perform a function (an input/output map) from data. Adaptive means that the system parameters are changed during operation, normally called the training phase. After the training phase the Artificial Neural Network parameters are fixed and the system is deployed to solve the problem at hand (the testing/validation phase). The Artificial Neural Network is built with a systematic step-by-step procedure to optimize a performance criterion or to follow some implicit internal constraint, which is commonly referred to as the learning rule. The input/output training data are fundamental in neural network technology, because they convey the necessary information to "discover" the optimal operating point. The nonlinear nature of the neural network processing elements (PEs) provides the system with lots of flexibility to achieve practically any desired output. There is a style in neural computation that is worth describing.

Basic Structure

An ANN is commonly divided into three or more layers: an input layer, a hidden layer(s), and an output layer. The input layer contains the input nodes (neurons), i.e. the input variables for the network. The output layer contains the desired output of the system and the hidden layer usually contains a series of nodes associated with transfer functions. Each layer of the ANNs is linked by weights that have to be determined through a learning algorithm. A simple three layer neural network is sown in Fig 1.2.

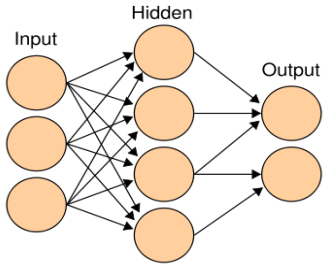

Figure 1. A three layer neural network

An input is presented to the neural network and a corresponding desired or target response set at the output. An error is composed from the difference between the desired response and the system output. This error information is fed back to the system and adjusts the system parameters in a systematic fashion (the learning rule). The process is repeated until the performance is acceptable.

It is clear from this description that the performance hinges heavily on the data. If one does not have data that cover a significant portion of the operating conditions or if they are noisy, then neural network technology is probably not the right solution. On the other hand, if there is plenty of data and the problem is poorly understood to derive an approximate model, then neural network technology is a good choice. ANN-based solutions are extremely efficient in terms of development time and resources, and in many difficult problems artificial neural networks provide performance that is difficult to match with other technologies. At present, artificial neural networks are emerging as the technology of choice for many applications, such as pattern recognition, prediction, system identification, and control.

Feed forward neural network

The feed forward neural network was the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. The data flow is strictly feed forward and there are no cycles or loops in the network. The data processing can extend over multiple (layers of) units, but no feedback connections are present, that is, connections extending from outputs of units to inputs of units in the same layer or previous layers. A three layer feed forward neural network with 3 inputs, 2 hidden nodes and 2 outputs is shown in Fig 1.4.

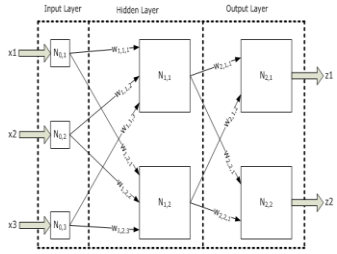

Figure 2. Feed Forward Neural Network

Backpropagation Algorithm

The Backpropagation algorithm is the most popular of the ANN training algorithms. The Backpropagation algorithm (BPA) is basically a procedure to train feed forward models (FFMs). It requires a teacher that knows, or can calculate, the desired output for any given input. The term is an abbreviation for "backwards propagation of errors". A three layer neural network with notations for backpropagation is shown in Fig 3.3.

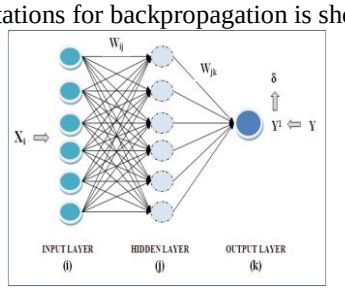

Figure 3. Three layer Network with Notations

A brief Back Propagation training algorithm is given below [8, 9].

1. Initialize the number of hidden nodes
2. Initialize the maximum number of iteration and the learning rate (η). Set all weights and thresholds to small random numbers. Thresholds are weights with corresponding inputs always equal to 1.
3. For each training vector (input $X_p = (x_1, x_2 ..... x_n)$, output Y) repeat steps 4-7.
4. Present the input vector to input nodes and the output to the output nodes;
5. Calculate the input to the hidden nodes:

$$a_j^h = \sum_{i=1}^{n} W_{ij}^h x_i$$

Calculate the output from the hidden nodes:

$$x_j^h = f(a_j^h) = 1/(1+\exp(-a_j^h))$$

Calculate the inputs to the output nodes:

$$a_k = \sum_{j=1}^{L} W_{jk} x_j^h$$

and the corresponding outputs:

$$\hat{Y}_k = f(a_k) = 1/(1+\exp(-a_k))$$

Notice that k = 1 and L is the number of hidden nodes.

6. Calculate the error term for the output node:

$$\delta_k = (Y - \hat{Y}) f'(a_k)$$

and for the hidden nodes:

$$\delta_j^h = f'\left(a_j^h\right)\sum_k \delta_k W_{jk}$$

7. Update weights on the output layer:

$$W_{jk}(t+1) = W_{jk}(t) + \eta\delta_k x_j^h$$

and the hidden layer:

$$W_{ij}(t+1) = W_{ij}(t) + \eta\delta_j^h x_i$$

While network errors are larger than some predefined limit or number of iteration is smaller than the maximum number of iterations repeat steps 4-7.

ANN in Modelling of Algal Blooms

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. ANNs can identify and learn correlated patterns between input data sets and corresponding target values. After training, ANNs can be used to predict the output of new independent input data. ANNs imitate the learning process of the animal brain and can process problems involving very nonlinear and complex data even if the data are imprecise and noisy. Thus they are ideally suited for the modelling of ecological data which are known to be very complex and often non-linear.

The causality and dynamics of algal production are extremely complicated and not well-understood. The capabilities of water quality and ecological models have greatly advanced in recent years, especially for predicting long term trends and when sufficient field data is available for model calibration and validation. Nevertheless, the prediction of algal primary production remains a very difficult problem. The uncertainty and the complexity of existing deterministic two- or three-dimensional models (often coupled with hydrodynamic models) have prevented their effective use as a forecasting tool [2]. An alternative approach aimed at more robust forecasts of primary production is therefore necessary.

To be able to control algal blooms, it is necessary to be able to determine the key factors governing the algal dynamics and to establish an algal response model which can effectively simulate the timing and magnitude of algal blooms. Recently, artificial neural network (ANN) technology has been applied in the prediction of algal blooms [12, 17].

In this framework artificial neural networks provide an effective alternative to conventional modeling techniques [13, 14] and their application to regional scale phytoplankton primary production modeling has already been presented [15].

Applications of ANN in ecological and environmental science have been reported since the beginning of the 1990s. The use of ANNs as ecological modelling tools has been the subject of a number of investigations [9]. The temporal changes of particular algal species in freshwater systems have been modelled [6, 12]. It is shown that ANNs are capable of simulating trends of algal growth dynamics.

Issues in ANN Modelling

ANNs offer a promising alternative approach to traditional linear methods. However, while ANNs provide a great deal of promises, they also embody a large degree of uncertainty. These problems are not easy to tackle. Some of the issues were discussed in this chapter.

Selection of Data Set

A training and a validation/test sample are typically required for building an ANN forecaster. The training sample is used for ANN model development and the validation sample is adopted for evaluating the forecasting ability of the model. Although there is no general procedure to select the data sets, several factors such as the problem characteristics, the data type and the size of the available data should be considered in making the decision. The literature offers little guidance in selecting the training and the test sample. Most authors select them based on the rule of 80% vs. 20%, 70% vs. 30%, 60% vs. 40%, etc. At least 20 percent of any sample should be used for validation process [18]. The amount of data for the network training depends on the network structure, the training method, and the complexity of forecast the particular problem. In general, as in any statistical approach, the sample size is closely related to the required accuracy of the problem. The larger the sample size, the more accurate the results will be. Even for less than two years of data set, ANN shows the better ability of capturing the pattern and the correlation coefficient values are well above 0.95 in many literatures [1, 13].

Network Architecture

An ANN is typically composed of layers of nodes. In the popular Feed Forward Network, all the input nodes are in one input layer, all the output nodes are in one output layer and the hidden nodes are distributed into one or more hidden layers in between. In designing an ANN, one must determine the following variables:

* The number of input nodes.
* The number of hidden layers and hidden nodes.
* The number of output nodes.

The selection of these parameters is basically problem dependent. To date, there is no simple clear-cut method for determination of these parameters. A three layer Feed Forward Neural Network is Shown in Fig 3.2.

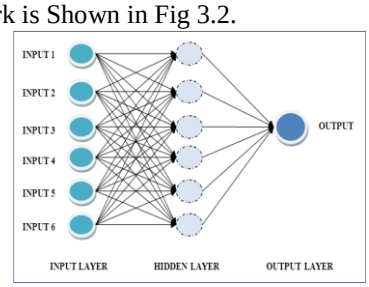

Figure 4. A three layer Feed Forward Neural Network

Number of Input Nodes

The number of input nodes corresponds to the number of input variables. However, currently there is no suggested systematic way to determine this number. Ideally, we desire a small number of essential variables which can unveil the unique features embedded in the data. Too few or too many inputs can affect either the learning or prediction capability of the network. The number of inputs is probably the most critical decision variable for any problem since it contains the important information about the complex (linear and/or nonlinear) structure in the data. The input variables are closely related to the desired output variables. Lee et al (2003), Yabunaka et al (1997), Recknagel et al (1997), Scardi and Harding (1999) and Scardi (2001) have studied the algal bloom characteristics with 10 input variables. Their investigations showed the correlation coefficient values more than 0.8. Similar correlation coefficients were obtained with less than 5 input variables in various studies [13, 15]. When the relationship to be modelled is not well understood, then an analytical technique, such as correlation analysis, is often employed to select inputs [11].

Number of Hidden Layers and Hidden Nodes

The hidden layer and nodes play very important roles for many successful applications of neural networks. It is the hidden nodes in the hidden layer that allow neural networks to detect

the feature, to capture the pattern in the data, and to perform complicated nonlinear mapping between input and output variables. Single hidden layer is sufficient for ANNs to approximate any complex nonlinear function with any desired accuracy. Most authors use only one hidden layer for modelling purposes [6, 7, 12, 16, 17]. Two hidden layer networks may provide more benefits for some type of problems. Several authors address this problem and consider more than one hidden layer (usually two hidden layers) in their network design processes. Srinivasan et al. (1994) use two hidden layers and this results in a more compact architecture which achieves a higher efficiency in the training process than one hidden layer networks. Zhang (1994) finds that a network never needs more than two hidden layers to solve most problems including modelling. In our view, one hidden layer may be enough for most forecasting problems. However, using two hidden layers may give better results for some specific problems, especially when one hidden layer network is over laden with too many hidden nodes to give satisfactory results. In general, networks with fewer hidden nodes are preferable as they usually have better generalization ability and less over fitting problem [18]. But networks with too few hidden nodes may not have enough power to model and learn the data. The most common way in determining the number of hidden nodes is via experiments or by trial-and-error. Networks with the number of hidden nodes being equal to the number of input nodes are also reported to have better forecasting results in several studies [18].

Number of Output Nodes

The number of output nodes is relatively easy to specify as it is directly related to the problem under study. From the previous studies it has been observed that for the modelling of algal blooms, the output nodes will contains the chlorophyll-a concentrations and some algal species with respect to the location. Recknagel et al (1997) studied a lake ecosystem with 10 output parameters contains different algal species responsible for the algal bloom. But in many studies, researchers have used only one output parameter (chlorophyll-a).

Interconnection between Nodes

The network architecture is also characterized by the interconnections of nodes in layers. The connections between nodes in a network fundamentally determine the behavior of the network. For most forecasting as well as other applications, the networks are fully connected in that all nodes in one layer are only fully connected to all nodes in the next higher layer except for the output layer. Adding direct links between input layer and output layer may be advantageous to forecast accuracy since they can be used to model the linear structure of the data and may increase the recognition power of the network. Also the connection carries some weightages from each node [18]. Initial weights will be randomly assumed and will be adjusted during the training process.

Activation Function

The activation function is also called the transfer function. It determines the relationship between inputs and outputs of a node and a network. In general, the activation function introduces a degree of nonlinearity that is valuable for most ANN applications. Any differentiable function can qualify as an activation function. In practice, only a small number of ‘‘well behaved’’ activation functions are used. These include,

1.The sigmoid (logistic) function:

$f(x) = (1+\exp(-x))^{-1}$

2.The hyperbolic tangent (tanh) function:

$f(x)=(\exp(x)-\exp(-x)) /(\exp(x)+\exp(-x))$

3.The sine or cosine function:

$f(x)=\sin(x)$ or $f(x)=\cos(x)$

4.The linear function:

$f(x)=x$

Where, x is the input to a node. Among them, logistic transfer function is the most popular choice. Generally, a network may have different activation functions for different nodes in the same or different layers. Yet almost all the networks use the same activation functions particularly for the nodes in the same layer [5, 17]. While the majority of researchers use logistic activation functions for hidden nodes, there is no consensus on which activation function should be used for output nodes. Following the convention, a number of authors simply use the logistic activation functions for all hidden and output nodes [6, 7, 12, 16, 17].

Training Algorithm

The neural network training is an unconstrained nonlinear minimization problem in which arc weights of a network are iteratively modified to minimize the overall mean or total squared error between the desired and actual output values for all output nodes over all input patterns. The most popularly used training method is the backpropagation algorithm [8, 9]. An error (□) is composed from the difference between the desired response and the system output. This error information is fed back to the system and adjusts the system parameters in a systematic fashion (the learning rule). The process is repeated until the performance is acceptable.

Data Normalization

Nonlinear activation functions such as the logistic function typically have the squashing role in restricting or squashing the possible output from a node to, typically, (0, 1) or (-1, 1). Data normalization is often performed before the training process begins. As mentioned earlier, when nonlinear transfer functions are used at the output nodes, the desired output values must be transformed to the range of the actual outputs of the network. The following types of data normalization techniques are frequently used.

1. linear transformation to [0,1]:

$xn=(x0-xmin) / (xmax-xmin)$

2. linear transformation to [a,b]:

$xn=(b-a)(x0-xmin)/(xmax-xmin) +a$

3. statistical normalization:

$xn=(x0-x)/s$

4. simple normalization:

$xn=x0/xmax$

Where xn and x0 represent the normalized and original data; xmin, xmax, x and s are the minimum, maximum, mean and standard deviations along the columns or rows respectively. Among all the four methods, linear transformation is widely used in many studies. The choice of range to which inputs and targets are normalized depends largely on the activation function of output nodes, with typically [0, 1] for logistic function and [-1, 1] for hyperbolic tangent function [18].

Performance Measures

Although there can be many performance measures for an ANN forecaster like the modeling time and training time, the ultimate and the most important measure of performance is the prediction accuracy it can achieve beyond the training data. An accuracy measure is often defined in terms of the forecasting error which is the difference between the actual (desired) and the predicted value. There are a number of measures of accuracy in

the forecasting literature and each has advantages and limitations [10]. The most frequently used are,

1.the mean absolute deviation (MAD)

MAD = et/

2. the sum of squared error (SSE)

SSE = (et)2

3. the mean squared error (MSE)

MSE = (et)2 / N

4. the root mean squared error (RMSE)

RMSE = (MSE)

5. the mean absolute percentage error (MAPE)

MAPE = (1/N)(et/yt)(100)

Where et is the individual forecast error; yt is the actual value; and N is the number of error terms. It is important to note that the first four of the above frequently used performance measures are absolute measures and MSE is the most frequently used accuracy measure in the literature.

Sensitivity Analysis

A sensitivity analysis will be performed to investigate the relationship between the output parameter and input water quality parameters. That is, in order to find out the effect of each input parameter on the output, this analysis will be performed [16, 17]. This will be done with the trained ANN after the validation process. This analysis will be performed by increasing the input values of a particular data set by 10% and the percent change in the output concentration over with original data set will be analyzed.

Conclusion

We have presented a review of the current state of the use of artificial neural networks for modelling application. The unique characteristics of ANNs, adaptability, nonlinearity and mapping ability, make them quite suitable and useful for modelling tasks. ANNs offer a promising alternative approach to traditional linear methods. Overall, ANNs give satisfactory performance in modelling. A considerable amount of research has been done in this area. The findings are inconclusive as to whether and when ANNs are better than classical methods. There are many factors that can affect the performance of ANNs. However, there are no systematic investigations of these issues. The shotgun (trial-and-error) methodology for specific problems is typically adopted by most researchers, which is the primary reason for inconsistencies in the literature. Also there are no structured methods today to identify what network structure can best approximate the function, mapping the inputs to outputs. Hence, the tedious experiments and trial-and-error procedures are often used. So the ANN modelling needs further more research for addressing these issues.

References

[1] Barciela R.M., Garcia E., Fernandez E., (1999), "Modelling primary production in a coastal embayment affected by upwelling using dynamic ecosystem models and artificial neural networks", Ecological Modelling. Vol 120, pp 199_/211.

[2] Delft, a project report (2000), "Use of artificial neural networks and fuzzy logic for integrated water management: Review of applications", Journal of Hydroinformatics.

[3] Dilon P.J. and Rigler F.H. (1974), "The phosphorous-chlorophyll relationship in lakes", Limnology and Oceanography. Vol 19, pp 767-773.

[4] Duarte P., Bernardo J.M., Costa A.M., Macedo F., Calado G.and Cancela L da Fonseca (2002), "Analysis of coastal lagoon metabolism as a basis for management", Aquatic Ecology. Vol 36, pp 3-19.

[5] Jan Tai Kuo, Ming Han Hsieh, Wu Seng Lung and Nian She (2007), "Using artificial neural network for reservoir eutrophication prediction", Ecological Modelling. Vol 200, pp 171-177.

[6] Karul C., Soyupak S., Cilesiz A.F., Akbay N., Germen E., (2000), "Case studies on the use of neural networks in eutrophication modelling", Ecological Modelling. Vol 134, pp 145–152.

[7] Lee J.H.W., Huang Y., Dickman M. and jayawardena A.W. (2003), "Neural network modelling of coastal algal bloom", Ecological Modelling. Vol 159, pp 179-201.

[8] Lek S., Delacoste M., Baran P., Dimopoulos I., Lauga J., Aulagnier S. (1996), "Application of neural networks to modelling nonlinear relationships in ecology", Ecological Modelling. Vol 90, pp 39–52.

[9] Lek S., Guegan J.F. (1999), "Artificial neural networks as a tool in ecological modeling, an introduction", Ecological Modelling. Vol 120, pp 65-73.

[10]Makridakis S., Wheelwright S.C., McGee V.E. (1983). Forecasting: Methods and Applications, 2nd ed. John Wiley, New York.

[11]Nitin Muttil and Kwok-Wing Chau (2007), "Machine-Learning paradigms for selecting ecologically significant input variables", Engineering Applications of Artificial Intelligent. Vol 20, pp 735-744.

[12]Recknagel F., French M., Harkonen P., Yabunaka K. (1997), "Artificial neural network approach for modelling and prediction of algal blooms", Ecological Modelling. Vol 96, pp 11–28.

[13]Scardi M. (1996), "Artificial neural networks as empirical models of phytoplankton production", Marine Ecology Progress Series. Vol 139, pp 289–299.

[14]Scardi M. (2001), "Advances in neural network modelling of phytoplankton primary production", Ecological Modelling. Vol 146, pp 33-45.

[15]Scardi M., Harding L.W. (1999), "Developing an empirical model of phytoplankton primary production: a neural network case study", Ecological Modelling. Vol 120, pp 213–223.

[16]Wei B., Sugiura N., Maekawa T. (2001), "Use of artificial neural network in the prediction of algal blooms", Water Research. Vol 35, pp 2022–2028.

[17]Yabunaka K., Hosomi M. and Murakami A. (1997), "Novel application of backpropagation artificial neural network model formulated to predict algal bloom", Water Science and Technology. Vol 36, pp 89-97.

[18]Zhang G., Patuwo B.E. and Hu M.Y. (1998), "Forecasting with artificial neural networks: The state of the art". International Journal of Forecasting. Vol 14, pp 35-62.