Journal of Engineering Science and Technology
Vol. 5, No. 3 (2010) 342 - 349
© School of Engineering, Taylor's University College

DATA MINING APPLICATION IN PREDICTING *CRYPTOSPORIDIUM SPP*. OOCYSTS AND *GIARDIA SPP*. CYSTS CONCENTRATIONS IN RIVERS

TOOCHUKWU C. OGWUELEKA[1,*], FRANCISCA N. OGWUELEKA[2]

[1]Department of Civil Engineering, University of Abuja, PMB 117, Abuja, FCT, Nigeria
[2]Department of Computer Science, University of Abuja, PMB 117, Abuja, FCT, Nigeria
*Corresponding Author: ogwueleka@yahoo.co.uk

Abstract

Data mining is a set of computer-assisted techniques designed to automatically mine large volumes of integrated data for new, hidden or unexpected information, or patterns. Two artificial neural networks (ANN) models were developed for prediction of *Cryptosporidium* oocysts and *Giardia* cysts respectively using multiple water quality parameters as input. These neural models were feed forward networks, trained by back propagation algorithm. Eight water quality parameters were used to predict *Cryptosporidium* peak concentration and seven parameters were used to model *Giardia* concentration in Kano River, Nigeria. The ANN models correctly predicted oocysts and cysts concentration with accuracy of 90% and 92% respectively. The neural network model gave excellent results.

Keywords: Artificial neural networks, Data mining, Water quality, Microbial contamination, Faecal pollution.

1. Introduction

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make a valid prediction [1]. Discovering previous unknown patterns, predicting new trends and behaviour are the two outstanding characteristics of data mining. The two primary goals of data mining tend to be prediction and description. Prediction involves using some variables or fields in the data set to predict unknown or future values of the other variables of interest and description focuses on finding patterns describing the data that can be interpreted by humans [2].

Nomenclatures

ANN	Artificial neural network
b_1^j, b_2^k	Bias terms
$f(x)$	Corresponding output at that node
$f_1(.)$, $f_2(.)$	Activation functions
K	Output nodes
L	Hidden nodes
N	Input nodes
O_{pk}	Output from the k^{th} node of the output layer of the network
R	Coefficient of correlation
$W_{i,j}^{h}$	Connection weight between i^{th} node of the hidden layer and j^{th} node of the output layer
$W_{j,k}^{o}$	Connection weight between j^{th} node of the hidden layer and k^{th} node of the output layer
x	A given input to a node in the neural network
x_{pi}	Inputs to the network for p^{th} vector
y_m	Actual value of concentration
y_s	Simulated value of concentration
$\bar{y}_m$	Average of actual value of concentration
$\bar{y}_s$	Average of simulated value of concentration

Several data mining techniques have been proposed to date such as decision tree, genetic algorithm, artificial neural network, nearest neighbour method, rule induction etc. These data mining techniques except artificial neural network cannot learn from new data and their precision is poor when inaccurate data are used.

An artificial neural network is a non-linear predictive model that learns through training and resembles biological neural networks in structure. ANNs as data driven empirical models have been successfully applied in all fields of water technology: metal bioleaching in municipal sludge [3], prediction of wastewater inflow rate [4], membrane technology modelling [5, 6], and to identify non-point sources of faecal contamination [7]. Basheer and Najjar [8] used ANN to predict the breakthrough time in a fixed bed adsorption system. Starrett et al. [9] employed an ANN to predict pesticide leaching through turf grass covered soil.

ANN has been used for the prediction of peak microbial concentrations; sorting land use associated faecal pollution sources and relative ages of runoff. Brion et al. [10] and Neelakantan et al. [11] carried out research on the ability of ANNs to predict peak microbial concentrations of *Giardia* and *Cryptosporidium* from multiple input parameters. ANN was used in this study because modelling applications for microbial water quality is difficult due to complexities in environmental distribution, mobility and fate of microbes. This study is limited to applications of ANNs to two separate microbial water quality databases.

The encysted waterborne parasites *Giardia* lamblia and *Cryptosporidium* parvum have presented a challenge to water suppliers since the 1980s [12]. One reason that encysted parasites have become an important drinking water issue is that they are the most frequently recognized as common causes of gastro-enteritis in both the developed and developing countries.

Waterborne parasites occur at varying concentration in drinking water sources [13, 14, 15]. The outbreaks have occurred as a result of a variety of conditions, including use of untreated surface water, contaminated water distribution systems, treatment deficiencies.

Cryptosporidium parvum is recognized as a frequent cause of waterborne disease in humans. *Cryptosporidium* parvum outbreaks from surface water supplies have been documented [16, 17]. The life cycle of *Cryptosporidium* parvum can be summarized by six events: excystation of the oocysts, in the intestine of the host, replication within the host, gamete formation, fertilization, oocyst wall formation and sporozoite formation. It has been observed that oocysts are capable of passing membrane filters greater than 1 μm in pore size. There are two stages in the life of *Giardia* lamblia in warm blooded vertebrates, the trophozoite and the cyst. Cyst is found in the small and large intestines and the faeces. The cyst is the normal infective stage in the transmission of Giardiasis to new hosts.

Neural Networks are a class of computational tools that operate analogously to the biological processes of a brain. Neural Network imitates the working of human neuron and works on stimulus from outside world. It consists of many single processors, which interact through a dense web of interconnections. A neuron or processing element (PE) has primarily two things to do. First, it computes output, which is sent to the other PE's or outside the network. The neuron or PE determines its output value by applying a transfer function. Secondly, it updates a local memory, that is, data variables. The neurons are organized into three generic types of layers. The first layer is called the input layer and the last layer is the output layer. The inner layers, one or more, are known as hidden layers. Input data are presented to the input layer and the passed to each hidden layer in sequence and finally to the output layer. Hidden neurons connect the input neurons to the output neurons and provide nonlinearity to the network. Figure 1 shows a schematic diagram of a multi-layer neural network.

The objective of this study is to design and develop an ANN model for prediction of peak *Cryptosporidium* and *Giardia* concentration in Kano River, Nigeria.

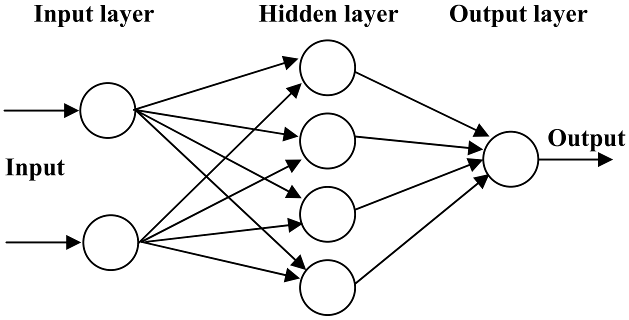

Fig. 1. A Sample of Neural Network Structure.

2. Approach and Methods

In this study, a six-step procedure was used in setting up the ANN as shown in Fig. 2. Firstly; the data to be used was defined and presented to the ANN as a pattern of input data with the desired outcome or target. Secondly, the data was categorized to be either in the training set or validation set. The ANN only uses the training set in its learning process in developing the model. The validation set is used to test the model for its predictive ability. Thirdly, the ANN structure was defined by selecting the number of hidden layers to be constructed and the number of neurons for each hidden layer. Fourthly, all the ANN parameters were set before starting the training process. Next, the training process was started. The training process involves the computation of the output from the input data and the weights. The back propagation algorithm was used to train the ANN by adjusting its weights to minimize the difference between the current ANN output and the desired output. Finally, validation process was used to test the predictions of the designed models on generalized data. Large number of ANNs was constructed with different structures and parameters before determining an acceptable model as there are no fixed rules in determining the ANN structure or parameter values.

Back propagation, or propagation of error, is a common method of teaching artificial neural networks how to perform a given task. The errors and the learning propagate backwards from the output nodes to the inner nodes. Back propagation was used to calculate the gradient of the error of the network with respect to the network's modifiable weights. This gradient was then used in a simple stochastic gradient descent algorithm to find weights that minimize the error.

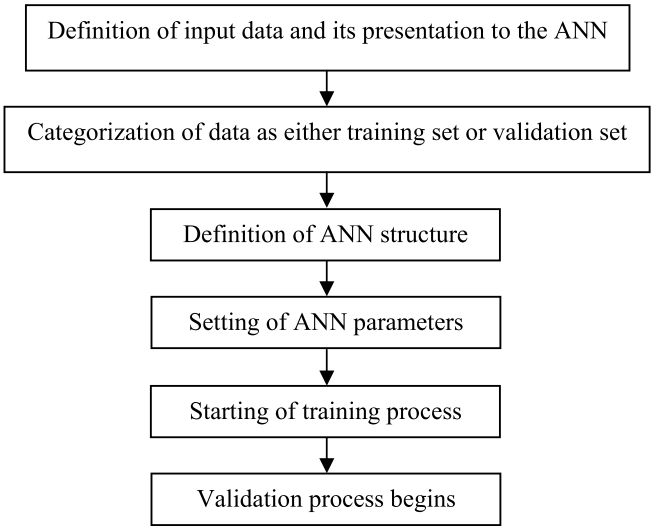
Definition of input data and its presentation to the ANN
Categorization of data as either training set or validation set
Definition of ANN structure
Setting of ANN parameters
Starting of training process
Validation process begins

Fig. 2. Six-step Procedure for Setting the ANN.

The ANN architecture used the structural design of N: 2N:1 where N was equal to the number of input parameters. The number of nodes in the hidden layer was set to twice that of the input layer as suggested by [7]. This type of network is called a feed forward network or multi-layer feed forward network. Feed forward networks are most commonly trained using a back-propagation algorithm. The three-layer, back propagation neural network can be expressed mathematically as

$$O_{pk} = f_1\left(\sum_{j=1}^{L} W_{jk}^{o} f_2\left(\sum_{j=1}^{N} W_{ij}^{h} x_{pi} + b_1^{j} \right) + b_2^{k} \right), \ \forall_k \in 1, 2, \ldots k \tag{1}$$

The commonly used activation function is a logistic sigmoidal function which has a form given below

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

where x is a given input to a node in the neural network and $f(x)$ is the corresponding output at that node. It is evident from Eq. (2) that the output from a neural network is always between 0 and 1.

Cryptosporidium and *Giardia* concentrations were modelled separately. Water quality parameters were obtained as input variables and peak concentrations were used as output. In training, according to Atherholt et al. [18], only eight water quality parameters were required to predict *Cryptosporidium* peak concentration by neural network. The eight parameters are pH, C. Perfinges, E.Coli., faecal coliform, turbidity, river flow, precipitation and total coliform. In case of *Giardia*, seven parameters were found to be essential (pH, perfinges, E. Coli., faecal coliform, turbidity, river flow, and alkalinity).

Six different sampling locations were taken along the Kano River at the inlet of Challawa drinking water plant in Kano, Nigeria. Six selected sites were sampled during the period of April 2008 to October 2008 with 60-database observation. The turbidity for each site was measured in the laboratory using turbidimeter. The pH was measured using standard pH meter. The rainfall data were obtained for meteorological agency in Kano, Nigeria. The river flow rates were obtained from Water Board in Kano. The parameters were analyzed as in Standard Methods for the Examination of Water and Wastewater [19].

Eight input parameters were used, amounting to sixteen hidden layers and a single output layer for modelling of *Cryptosporidium* peak concentration. Seven input parameters were used, amounting to fourteen hidden layers and a single output layer for modelling of *Giardia*. The steps of back propagation network are found in [20, 21]. The algorithmic procedure used in training the network is:

1. Apply input observations from training set to the network
2. Calculate corresponding output values
3. Compare the computed output with the known output values
4. Determine a measure of the error
5. Determine corrections, either as increase or decrease to the connection weights
6. Apply the corrections to the weights.

Repeat step 1-6 with all the training vectors until the error for all vectors in the training set is reduced to an acceptable value.

3. Results and Discussion

ANN was trained and simulated using code in Visual Basic. Net developed in house in University of Abuja, Nigeria. Performances of all ANN models were measured by coefficient of correlation, *R*, given by (3).

$$R = \frac{\sum_{i=1}^{n}\left(y_m - \bar{y}_m\right)\left(y_s - \bar{y}_s\right)}{\sqrt{\sum_{i=1}^{n}\left(y_m - \bar{y}_m\right)^2\left(y_s - \bar{y}_s\right)^2}}\,,\ \forall_k \in 1, 2, \ldots k \qquad (3)$$

The sixty dataset was grouped into two different sets each with the randomly chosen respective sampling points for training and testing of the ANN models. In predicting the peak concentrations, the modelling results obtained from training and validation of the two ANNs on sixty database observation assembled from measurements are shown in Table 1. After training and then during testing, it was observed that the ANN models were able to predict when an observation would have peak concentrations of encysted protozoa. In this study, peak is defined as above the 75th percentile of total cysts and full cysts recovered from an equivalent 5 litres sample.

Thirty seven training dataset were used for training. The inputs were presented to the input nodes and the final values were obtained after training. The obtained results showed that for *Cryptosporidium*, using input parameters, the trained ANN model correctly predicted oocyst concentration categories of peak or non-peak with 95 % accuracy. It was also noted that the trained ANN did not misclassify peak concentrations as non-peak, that is, false negative for the validation set of observations. Using input parameters, for *Giardia* with a peak concentration as three cysts with at least one cyst showing contents, the trained ANN model correctly predicted cyst concentration categories (peak or non-peak) with 97 % accuracy. The prediction values obtained were closely clustered around 0 and 1, which is an indication that the ANN was performing a linear cluster analysis from multiple input values. The clustering effect showed that the relationship the ANN was learning was clear and is related to the underlying distribution of encysted protozoa in the environment. These factors contributed to the mobility control.

Twenty three validation dataset were used for ANN models testing. For *Cryptosporidium*, when the simulated concentrations were compared with the actual concentrations, it yielded an accuracy of 90%. For *Giardia*, simulated concentrations were distinguished from actual concentration with the accuracy of 92%.

Table 1. ANN Prediction of Peak and Non-peak Encysted Protozoa Events.

Organism	ANN architectural design	Peak Concentration	*R* (Training)	*R* (Testing)
Cryptosporidium	8:16:1	≥2 full oocyst	0.95	0.90
Giardia	7:14:1	≥5 cycts with ≥1 full	0.97	0.92

4. Conclusions

ANNs were successfully used to predict the *Cryptosporidium* and *Giardia* peak concentrations in Kano River. Multilayer feed forward networks were used. In predicting *Cryptosporidium* peak concentration, the network had eight water quality input variables, sixteen neurons in the hidden layer and a single output layer. For *Giardia* peak concentration seven input variables were used. It was observed that ANNs provided a linear and cluster analysis from various parameter data observations. The research has demonstrated that a neural network can be used to predict microbial concentration from water quality data, and that the technique can be automated.

References

1. Edelstien, H.A. (1999). *Introduction to data mining and knowledge discovery* (3rd Ed.). Two Crows Corporation.
2. Bolton, R.J; and Hand, D.J. (2002). Unsupervised profiling methods for fraud detection*. Statistical Science*, 17(3), 235-255.
3. Laberge, C.; Cluis, D.; and Mercier, G. (2000). Metal bioleaching prediction in continuous processing of municipal sewage with *Thiobacillus Ferrooxidans* using neural networks. *Water Research*, 34(4), 1145-1156.
4. El-Din, A.G.; and Smith, D.W. (2002). A neural network model to predict the wastewater inflow incorporating rainfall events. *Water Research*, 36(5), 1115-1126.
5. Strugholtz, S.; Panglisch, S.; Gebhardt, J.; and Gimbel, R. (2008). Neural networks and genetic algorithms in membrane technology modelling. *Journal of Water Supply: Research and Technology-AQUA,* 57 (1), 23-34.
6. Cabassud, M.; Delgrange-Vincent, N.; Cabassud, C.; Durand-Bourlier, L.; and Laine, J.M. (2002). Neural network: a tool to improve UF plant productivity. *Desalination*. 145 (1-3), 223-231.
7. Brion, G.M.; and Lingireddy, S. (1999). A neural network approach to identifyimg non-point sources of microbial contamination. *Water Research*, 33(14), 3099–3106.
8. Basheer, I.A.; and Najjar, Y.M. (1995). Designing and analyzing fixed bed adsorption systems with artificial neural networks. *Journal of Environmental Systems*, 23(3), 291–312.
9. Starrett, S.K.; Najjar, Y.M.; and Hill, J.C. (1996). Neural networks predict pesticide leaching. *Proceedings of the American Water and Environment Conference*. New York: ASCE, 1693–8.
10. Brion, G.M.; Neelakantan, T.R.; and Lingireddy, S. (2001). Using neural networks to predict peak *Cryptosporidium* concentrations. *Journal of the American Water Works Association,* 93(1), 99–105.
11. Neelakantan, T.R.; Lingireddy, S.; and Brion, G.M. (2002). Relative performance of different ANN training algorithms in predicting protozoa concentration in surface waters. *ASCE Journal of Environmental Engineering*, 128 (6), 533–542.

12. Finch, G.R.; and Belosevie, M. (2002). Controlling *Giardia ssp.* and *Cryptosporium spp.* in drinking water by microbial reduction processes. *Journal of Environmental Engineering and Science*, 1, 17-31.
13. Smith, H.V.; and Rose, J.B. (1998). Waterborne *Cryptosporidiosis*: current status. *Parasitology Today*, 14(1), 14-22.
14. Karanis, P.; Schoenen, D.; and Seitz, H.M. (1998). Distribution and removal of *Giardia* and *Cryptosporidium* in water supplies in Germany. *Water Science and Technology*, 37(2), 9-18.
15. Zuckerman, U.; Gold, D.; Shelef, G.; and and Armon, R. (1997). Presence of *Giardia* and *Cryptosporidium* in surface waters and effluents in Israel. *Water Science and Technology*, 35(11-12), 381-384.
16. Craun, G.F. (1991). Causes of waterborne outbreaks in the United States. *Water Science and Technology,* 24(2), 17-20.
17. Poulton, M.; Colbourne, J.; and Dennis, P.J. (1991). Thames water's experiences with *Cryptosporidium*. *Water Science and Technology*, 24(2), 21-26.
18. Atherholt, T.B.; LeChevallier, M.W.; Norton, W.D.; and Rosen, J.S. (1998). Effect of rainfall on *Giardia* and *Cryptosporidium*. *Journal of the American Water Works Association*, 90(9), 66-80.
19. Standard Methods (1995). *Standard methods for the examination of water and wastewater* (17th and 19th Eds). Washington, DC: APHA, AWWA and WEF.
20. Freeman, J.A.; and Skapura, D.M. (1991). *Neural networks: algorithms, applications and programming techniques.* Addison-Wesley Pub (Sd).
21. Masters T. (1993). *Practical neural network recipes in C++.* Academic Press, USA.