Features Selection in Water Quality Prediction in Neural Network using Canonical Correspondence Analysis (CCA)

N. A. Rahim, Z. Ahmad

Abstract

Irrelevant inputs can cause deterioration of the network performance. This paper aims to implement features selection for water quality prediction model as the neural network generalization may be improved when the number of measured variables is reduced. Canonical Correspondence Analysis (CCA) was introduced as a feature selection method, where it choose a subset of input variables by eliminating features with little or no predictive information. Data monitoring and sampling process was carried out at 5 sampling sites of Perak river basin. The water quality data with 28 different parameters was used to determine the relative effect for each parameter to the biological oxygen demand (BOD) and chemical oxygen demand (COD) as they represent the river water quality; determination of the pollutants strength in terms of the oxygen necessity to reduce the domestic and industrial wastes, hence giving an initial value how much biodegradable waste existed in the water. While, for COD, its application is practical since it verify the amount of total oxidizable compounds in water. Results of CCA through the biplot diagram showed that suspended solid (SS), turbidity (TUR), total solid (TS), nitrate (NO₃) and zinc (Zn) were most important environmental factors influencing the BOD and COD. The study continued with a number of neural network approaches implemented for predicting BOD and COD. The results showed comparable performance of water quality prediction with CCA implementation in the network compared to ANN standalone model.

Keywords: Artificial Neural Network; Canonical Correspondence Analysis; Feature selection; Water quality index; Water quality prediction

1. Introduction

River monitoring is a process whereby important elements of a river is measured in order to assess and sustain or improve the health of a river and its ecosystem. In other words, that activity is actually determining the river water quality which is a phrase to describe the chemical, biological and physical characteristic of water. They add up to give a full image of the state of a river. In this regard, Malaysian Government bodies that are responsible for the management of rivers in Malaysia are the Department of Environment (DOE) and the Department of Irrigation and Drainage (DID) with the cooperation of private sectors such as Alam Sekitar Malaysia Sdn. Bhd. (ASMA) (Juahir, et al., 2004).

In the case of artificial neural networks (ANNs), and other similarly data-driven statistical modeling approaches, there is no assumption made with respect to the structure of the model. Instead, the input variables are selected from the available data, and the
model is developed subsequently. R. May et al. (2011) mentioned that the difficulty of selecting input variables arises due to (i) the number of available variables, which may be very large; (ii) correlations between potential input variables, which create redundancy; and (iii) variables that have little or no predictive power (Suzuki, 2011; Talib et al.). In the process development of the model, input parameters must be chosen, based on the proven relationship to the outputs. According to Tarassenko (1998), the condition of input data somehow may affect the network; thus reducing the input dimensionality is a must by limiting the number of free parameters exists in the network. The free parameters are the total number of biases and weights within the network and are determined mainly by the number of inputs. Therefore, selection the correct input for the corresponding output prediction is so important in neural network modeling. Furthermore, the main reason of introducing input selection or input feature selection to the prediction model is to improve the prediction performance of the predictors and also providing faster and more cost-effective predictors or modeling.

Implementing Canonical Correspondence Analysis (CCA) as feature selection is popular among ecologists, which is this methods deals with an analysis of a correlation of variance-covariance matrix of a multivariate quantitative data set. The CCA can be thought of as a technique to analyze two sets of data, one in the form of a contingency table of species abundance at different sites and the other of external environmental variables (quantitative) on the same site. One objective of CCA is to develop relationships between these two different types of data. The output for CCA has several interesting parts (algebraically, numerically and graphically) that require interpretations to end users. The climax of this program is about constructing a biplot of the A matrix.

The aim of this study is to investigate the most descriptive variables to predict water quality using Canonical Correspondence Analysis (CCA) for the input of feedforward neural network model (ANN). The CCA approach allows the user to obtain a simultaneous representation of the sites, the objects, and the variables describing the sites in two or three dimensions that are optimal for a variance criterion.

2. Methodology

2.1 Sampling Methods

The study area is situated in the state of Perak, where there are 11 major river basins that covers over 80 kilometres squares. The purpose of the case study is to predict the water quality of the Perak River basin. The data were collected from March 2000 until November 2004, by Department of Environment of Malaysia. There are 30 water quality parameters available including physical and chemical parameters that were completely monitored from 48 monitoring stations in Perak river basin. A total 1008 samples of data were used for the canonical correspondence analysis. Real environmental data with noise during the monitoring process will give a good evaluation for the developed prediction model.

2.2 Data Analysis

The sampling data were normalized before analyzed with the CCA method. The normalization of the sampling data was needed due to the differences in unit of measurement during the data monitoring and recording process. The normalization process was done using MATLAB 7.11.0 (r2010b), using the predefined and built-in normalization function.

Further, the normalized data were analyzed with CCA method using XLSTAT 2012 software. CCA is one of the main uses of correspondence analysis in ecology, which visualizes a matrix of biological/environmental data in relation to a set of concomitant environmental variables. The canonical eigenvalues and the significance of
the relationships between the parameters and the canonical axes are tested by Monte Carlo permutations (p<0.05), using 499 permutations as implemented in XLSTAT 2012.

The constrained ordination diagram of CCA was carried out to define the best explanatory water quality parameters to the BOD and COD. Then, the results of the analysis were tested using ANN model whereby the model performance was evaluated using regression value (r-value) and mean square error (MSE). R-value provides the inconsistency measure of the data reproduced and fitness to the model while MSE evaluated the residual error between actual values and predicted values (Abdul Zali, 2011). The smaller values ensure the better performance (Rankovic et al., 2010).

2.2 Artificial Neural Network Prediction Model
Artificial neural network models were specified by the network topology, training and/or learning rules. These aspects have primarily affected the network performance; with three different layers in the network topology can be distinguished as: 1) An input layer: connecting the input information to the network. In this study, two sets of input nodes were applied; 5 and 28; 2) Hidden layer: acting as an intermediate computational layer. The numbers of hidden neuron were ranging between 15 to 40; 3) Output layer: producing the desired outputs. Two output nodes were utilised for the ANN prediction model; COD and BOD.

The ANN architecture for water quality prediction were decided by the performance of the respecting networks. The training algorithm for the whole networks utilizes the Levenberg-Marquardt Back Propagation algorithm, while the other neural network parameters used were; log-sigmoidal (logsig) and linear transfer function (pureline) as training function, 0.05 for learning rate, 1000 for epoch and 0.001 for goal.

The ANNs were trained for both type of inputs and different number of hidden neurons, sequentially selected to find the best performing network in the training data. An early stopping approach also applied to the network training. The best performance network on the training data was selected as trained network, and featured in the water quality predictor network.

3. Results and Discussion
3.1 Canonical Correspondence Analysis
The focal and significance point of this research was in fact, to illustrate in some extent the relation and distribution pattern of the environmental parameters to the water quality using CCA approach. The results of CCA are shown in Figure 1, where the empty square on the ordination diagram represent the individual environmental parameters for water quality.

The relative importance of a single parameter is depicted in the CCA ordination diagram by the length of their corresponding lines. For instance, Zn and SS had the longest lines and therefore showed its best correlation with COD and BOD (water quality) (Bodaghabadi et al., 2011). Meanwhile, the angles between the lines showed the interrelated level between the environmental parameters. The closer angle shows the high interrelation of those particular parameters. For example, the Tur and SS were interrelated according to Figure 1.

However, with canonical correspondence analysis and correlation coefficient both were considered for the feature selection, the environmental parameters that best related to the COD and BOD were Zn, SS, TS, Tur and NO3. The selected environmental parameters in CCA were used as new input data set for neural network
prediction model, while all environmental parameters as the default input data set for the neural network prediction model.

![CCA ordination of the 28 environmental parameters (empty square).](image)

Figure 1. CCA ordination of the 28 environmental parameters (empty square).

3.2 Artificial Neural Network Prediction Model

Based on the CCA method, only five environmental variables (TS, SS, ZN, NO₃ and Tur) were selected as new input data set to the CCA-ANN prediction model. For comparison, full input which contains DO, SS, pH, PB, CR, CD, Cond, Sal, Tur, DS, TS, Hg, NO₃, Cl, PO₄, As, Zn, OG, Ca, Fe, K, Mg, Na, NH₃-NL, MBAS, E-coli, Coliform and Temp were used as input for ANN prediction model. The justification of the ANN prediction models were based on these two sets of input data. Two sets of data, CCA-ANN and ANN were applied to the prediction models for training. The performance of each model in predicting the water quality was summarized in Figure 2 and Figure 3.

![The ANN prediction model performances (MSE) of COD for different number of hidden neurons for ANN and CCA-ANN input data sets.](image)

Figure 2. The ANN prediction model performances (MSE) of COD for different number of hidden neurons for ANN and CCA-ANN input data sets.
Figure 2 shows the ANN prediction model performance from validation data in terms of mean square error (MSE) using CCA-ANN and ANN input data set for different number of hidden neurons for COD output. In that figure, ANN prediction model shows the best performance for COD outputs using 36 hidden neurons for ANN input model and 12 hidden neurons for ANN-CCA input model. The similar performances were shown by the BOD output model. Their respective best performances were summarized as in Table 1.

Table 1. Best performances on each prediction model

<table>
<thead>
<tr>
<th></th>
<th>Performances on BOD</th>
<th>Performances on COD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA-ANN model</td>
<td>0.3649</td>
<td>0.1081</td>
</tr>
<tr>
<td>ANN model</td>
<td>0.3475</td>
<td>0.0593</td>
</tr>
</tbody>
</table>

Table 1 shows that both of the prediction models have a good prediction and best performance with a comparable score of MSE between the models. The results proved that the reduce-input parameters for the CCA-ANN prediction model was possible to delivered nearly same information as the full input parameters prediction model.

Figure 3 shows the r-value from validation data for different number of hidden neurons and outputs for both CCA-ANN and ANN prediction model for COD output. The network performance was evaluated based on the value of r (regression) to 1, with any value nearly to 1 was considered as good performance. COD output was rightfully predicted by CCA-ANN and ANN prediction model using 28 and 18 hidden neurons, respectively. The summarization in Table 2 shows the best r-value for each prediction model, as the similar result obtained for the BOD output model.

Figure 3. The ANN prediction model r-value of COD for different number of hidden neurons for ANN and CCA-ANN input data sets.
Both CCA-ANN and ANN prediction model have their best r-value around 0.9800 to 0.9900, indicating that both of the models can produce a very good prediction and performances. Their comparable score also proved that the CCA-reduced input model can be implemented in the ANN prediction model; with neural network generalization also could be improved. This result also indicates that the ANN also agree with CCA input selection to represent overall information in the data collection.

**Conclusion**

This study clearly shows that it is possible to reduce the number of the environmental parameters needed to make the water quality prediction, without any losses of any information and still have a comparable performance. From the ANN performance analysis and observation, the performance and ANN generalization also was improved with the reduce number of input parameters, help in reducing the time needed for network training and neuron computing time.

**Acknowledgment**

The authors are thankful to the Ministry of Higher Education Malaysia for sponsoring the research study fees and special gratitude to Department of Environmental (DOE) Malaysia for providing the river water quality data for this research.

**References**

A. Talib, Y. Abu Hasan and N. N. Abdul Rahman, in 18th World IMACS / MODSIM Congress, Cairns, Australia.
R. May, G. Dandy and H. Maier (2011), Review of Input Variable Selection Methods for Artificial Neural Networks, Artificial Neural Networks - Methodological Advances and Biomedical Applications, Prof. Kenji Suzuki (Ed.), InTech Open,