EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING

(APPLIED DATA SCIENCE)

IBM PROJECT-TEAM ID : PNT2022TMID39608

TEAM LEAD

SUDHARSAN S 510619104066

TEAM MEMBERS

SIDDHARTH N 510619104064

SUHAIL F 510619104068

VIGNESH M 510619104076

Of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

C.ABDUL HAKEEM COLLEGE OF ENGINEERING AND TECHNOLOGY

ANNA UNIVERSITY - CHENNAI 600 025

LITERATURE REVIEW

Survey 1:

ALI, M. QAMAR, A.M. (2013):

DATA ANALYSIS, QUALITY INDEXING AND PREDICTION OF WATER QUALITY:

Ali and Qamar et al. used the unsupervised technique of the average linkage (within groups) method of hierarchical clustering to classify samples into water quality classes. However, they ignored the major parameters associated with WQI during the learning process and they did not use any standardized water quality index to evaluate their predictions.

Survey 2:

X.SUN ,Y.LI (2013):

PREDICTING DISSOLVED OXYGEN FLUCTUATIONS IN GOLDEN HORN BY FUZZY TIME SERIES:

A water quality data is a kind of time series dataset which is likely to have complicated linear and non linear relationships.

The Fuzzy time series (FTS) model was first proposed by Song et al. and Chissom et al. in 1993 to address an enrolment prediction problem.

Chen et al. improved this model by replacing complicated max-min composition operations with simplified arithmetic operations. A Heuristic Gaussian cloud transformation was integrated with an FTS model to forecast water quality.

The experimental results showed that their proposed model significantly improved the prediction accuracy. However, there were only 520 water quality samples available to build the cloud, and thus, the model was not reliable or robust.

Time series analysis is also proposed to address dissolve oxygen prediction, and the experimental results show that the proposed analysis method can find out valuable knowledge from water quality historical time series data.

This algorithm has the smallest error in training and testing stages for the parameter Dissolved oxygen.

Survey 3:

A.H.ZARE (2014):

PREDICTION OF WATER QUALITY INDEX USING MULTIVARIATE LINEAR REGRESSION (MLR):

It is a kind of statistical analysis method that estimates the target values based on set of independent variables .This model is used to measure the Biological Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) based on the following four factors namely temperature, pH, total suspended solids and suspended solids.

This approach proposed that deterministic and multivariate linear regression models were used to speed up the process of predicting the water quality but as the dataset is considered as time series based, so it is likely to have a non-linear relationship. So, the performance of this algorithm is expected to be poor, with large prediction error.

Survey 4:

A.SARKAR AND P. PANDEY (2015):

RIVER WATER QUALITY MODELLING USING ARTIFICIAL NEURAL NETWORK TECHNIQUE:

This approach proposes that a time series prediction model was integrated with the ANN model to improve the prediction performance. A comprehensive comparision between ANN and MLR models in Biological Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) prediction has shown that the ANN model is a better option.

The major disadvantage in this proposed model is that the input parameters are ambigious and neural network struggle to formulate a non linear relationship in some scenarios.

Survey 5:

A. A. M. AHMED AND S. M. A. SHAH (2017):

APPLICATION ADAPTIVE NEURO FIZZY INFERENCE SYSTEM (ANFIS) TO ESTIMATE THE BIOLOGICAL OXYGEN DEMAND (BOD) AND CHEMICAL OXYGEN DEMAND (COD):

Many studies have proven that ANFIS, which can integrate linear and non-linear relationships hidden in the dataset, is a better option in this scenario.

This proposed approach is also used in predicting the effluent water quality and also shows that the ANFIS model works better than the ANN model in predicting the Dissolved Oxygen content in the water sample to be tested.

Even though there are only 45 data samples available, an ANFIS model with eight input parameters is used to predict total phosphorus and total nitrogen, the experiment result based on 120 water samples shows the proposed model is reliable. The ANFIS model has also been applied to estimate the biochemical oxygen demand in the Surma River. The testing results from 36 water samples confirmed that the ANFIS model could accurately formulate the hidden relationship and correlation analysis can improve the prediction accuracy.

The disadvantage in this proposed models shows that this approach requires that the size of the training dataset should not be less than the number of training parameters and if the correlation between the data in the dataset are weak then it generates out of range errors.

It shows poor performance in testing because the limited training dataset cannot build a robust or reliable model.

Survey 6:

AHMAD, Z.; RAHIM, N.; BAHADORI, A.; ZHANG, J. (2017):

IMPROVING WATER QUALITY INDEX PREDICTION THROUGH A COMBINATION OF MULTIPLE NEURAL NETWORKS:

Ahmad et al. employed single feed forward neural networks and a combination of multiple neural networks to estimate the WQI. They used 25 water quality parameters as the input. Using a combination of backward elimination and forward selection selective combination methods, they achieved an R2 and MSE of 0.9270, 0.9390 and 0.1200, 0.1158, respectively. The use of 25 parameters makes their solution a little immoderate in terms of an inexpensive real time system, given the price of the parameter sensors.

Survey 7:

SHAFI, U.; MUMTAZ, R.; ANWAR, H.; QAMAR, A.M.; KHURSHID, H. (2018):

SURFACE WATER POLLUTION DETECTION USING INTERNET OF THINGS:

Shafi et al. estimated water quality using classical machine
learning algorithms namely, Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks (Deep NN) and k Nearest Neighbors (kNN), with the highest accuracy of 93% with Deep NN.

The estimated water quality in their work is based on only three parameters: turbidity, temperature and pH, which are tested according to World Health Organization (WHO) standards. Using only three parameters and comparing them to standardized values is quite a limitation when predicting water quality.

REFERENCES

1. Ali, M. Qamar, A.M. Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in Pakistan. In Proceedings of the Eighth International Conference on Digital Information Management (ICDIM 2013), Islamabad, Pakistan, 10–12 September 2013; pp. 108–113.

2. X. Sun ,Y. Li "Predicting Dissolved Oxygen Fluctuations In Golden Horn By Fuzzy Time Series" International Journal of Nonlinear Science, vol. 17, no.3, pp. 234-240, Dec. 2013

3. A. H. Zare, "Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters," Journal of Environmental Health Science & Engineering, vol. 12, no. 1, pp. 1-8, Jan. 2014.

4. A. Sarkar and P. Pandey, "River Water Quality Modelling Using Artificial Neural Network Technique," Aquatic Procedia, vol. 4, pp. 1070-1077, 2015

5. A. A. M. Ahmed and S. M. A. Shah, "Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River," Journal of King Saud University - Engineering Sciences, vol. 29, no. 3, pp. 237-243, Jul. 2017.

6. Ahmad, Z.; Rahim, N.; Bahadori, A.; Zhang, J. Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. Int. J. River Basin Manag. **2017**, 15, 79–87.

7. Shafi, U.; Mumtaz, R.; Anwar, H.; Qamar, A.M.; Khurshid, H. Surface Water Pollution Detection using Internet of Things. In Proceedings of the 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), Islamabad, Pakistan, 8–10 October 2018; pp. 92–96.