Literature Survey on
Efficient Water Quality Analysis & Prediction Using Machine Learning

Team Members:

- PRABAKARAN S
- PERUMAL K
- KAVIYA N
- RAJA P

Team IBM Link:

https://github.com/IBM-EPBL/IBM-Project-3498-1658570711

STATEMENT:

- World is being surrounded by 3/4th of water surface and it is essential for all humans and living organisms.
- Quality of water is unstable, water can be polluted at any time and water quality testing is so expensive and a huge wastage of water.
- With that initiative we started a machine learning algorithm to estimate the quality of water.
-

Abstract

In this project, we aim to design a web-based monitoring system for efficient water quality analysis using machine learning. The current and present water quality data will be visualized by our web-based system. With our project, the quality of water will be predictable using machine learning algorithms. There are two parts to this project. Analyzing the water quality data for rivers, lakes, seas and analyzing the data for water treatment plants. The data for water treatment plants include samples taken from both the inlets and outlets of these plants.

Introduction

Water is the most important source of life. Water covers 71% of the earth and only 3% of water is fresh. Humankind always settled down near freshwater sources. Water has an enormous effect on human life throughout history. Even in ancient times, people found ways to purify water or to keep it clean. Considering people drink about 2-3 liters of water per day, they need to be sure that the water is clean and drinkable. Water is the home for the microorganism if there are no toxic chemicals in it. Although most microorganisms are harmless, there can be viruses or bacteria that can cause health damage. Also, there are toxic inorganic matters that can not be tolerated.

The effect of climate change and increasing demand for water by rapidly increasing population, industrialization, agricultural and other sectors is putting serious pressure on quality and quantity of water resources. For those reasons, managing and monitoring water, and detecting potential dangers before they affect water is highly important for protection and cleansing of water resources.

Therefore, The Ministry of Environment and Urban Planning is monitoring physical, chemical and biological parameters of important rivers, lakes, drainage channels and seas inside the Special Environmental Protection Area (SEPA).

There are 19 wastewater plants which analyze samples taken from 254 SEPA spots. The data gathered are stored in a database.

Managers and analysts need operational tools that help understanding the complex information about quality of water. Tools based on statistical approaches are often unable to conduct a detailed analysis due to sparse data and the invisible interactions of analysis results. This module will include basic data management functions such as time series management, spatial selection and representation, data availability assessment and data series comparison (visual and statistical).

Improved water quality prediction, accuracy and reduced computational complexity are vital for precise control over water quality. For this purpose, our aim is to develop a machine learning model that effectively predicts water quality and establishes an early warning system for water pollution.

Related Work

Machine Learning Methods for Better Water Quality Prediction
In this research paper, the dataset was created with 4 monitoring states on Johor River, a river in Johor State in Malaysia. A comparison is made between the following machine learning algorithms: WDT-ANFIS, ANFIS, RBF-ANN, and MLP-ANN. Due to the presence of noise in the data, it is relatively difficult to make an accurate prediction. Hence, a Neuro-Fuzzy Inference System based

augmented wavelet de-noising technique has been recommended that depends on historical data of the water quality parameter.

Dataset and Data Processing

Selecting the input variables for a model is very important for Artificial Neural Networks. The following water quality parameters were chosen for ANN modeling: temperature, electrical conductivity, salinity, nitrate(NO3), turbidity, phosphate(PO4), chloride(Cl), potassium(K), sodium(Na), magnesium(Mg), iron(Fe) and Escherichia coli(E-coli). These input parameters were used in many previous studies for ANN models. Using these parameters the prediction of pH, suspended solids (SS), and ammoniacal nitrogen (AN) is made possible.

Quality Performance

There are in total 3 models for three primary water quality parameters: AN, SS, and pH. The performance of the models is measured with the Coefficient of Efficiency (CE). Mean Square Error (MSE) is used to see the level of fitness between the network output and the desired output. Performance is better with smaller MSE values. Coefficient of Correlation (CC) is employed to inspect the linear relationship between the measured and predicted dissolved oxygen in the water. Using this methodology, the WDT-ANFIS models outperformed others.

Study of Short-Term Water Quality Prediction Model Based on Wavelet Neural Network

This research paper combines the wavelet transform with a Back Propagation (BP) neural network to build a short- term water prediction model. The trained model is used to predict the water quality on freshwater pearl breeding ponds in Duchang County, Jiangxi province, China. Also, a comparison has been made between Elman Neural Network, Wavelet Neural Network (WNN), and a BP network. The proposed model also features a high learning speed and improved accuracy.

Model Performance

Model performance was measured by Absolute Percentage Error (APE) and Mean Absolute Percentage Error(MAPE). The Wavelet Neural Network (WNN) outperformed BPNN and Elman NN by significantly lower APE. The model accuracy was greater than 90%. As shown in Figure 1, WNN also has higher prediction precision, stronger learning, and generalization ability compared to BPNN and Elman NN.

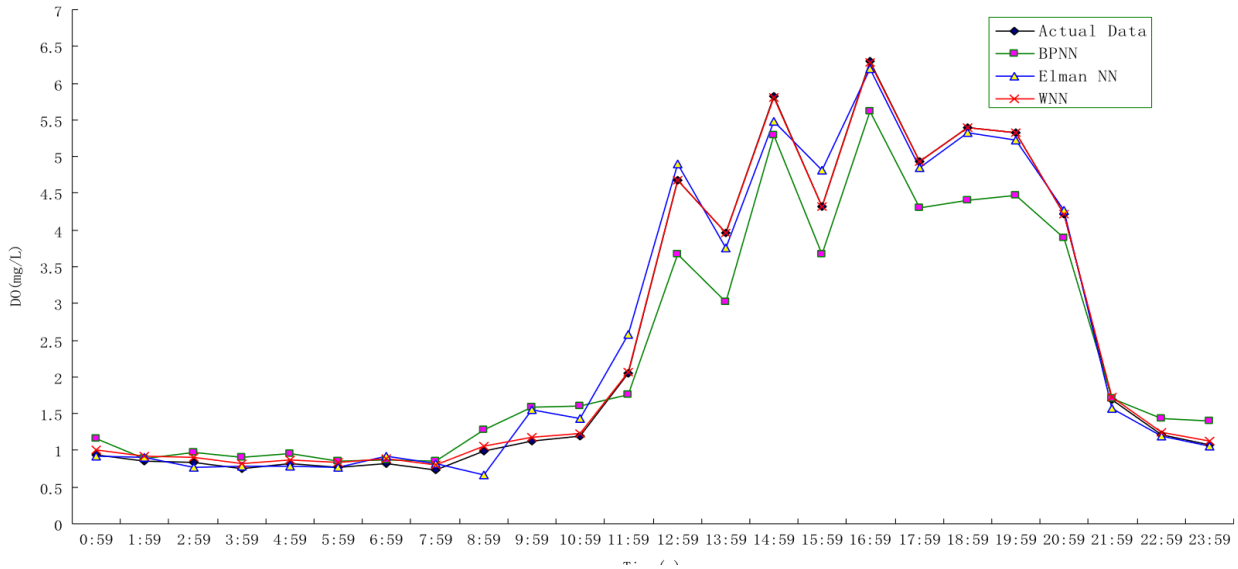

Figure 1 Figure 1: WNN compared to BPNN, Elman NN, and Actual Data. (x axis: time in minutes and seconds, y axis: Predicted dissolved oxygen mg/L)

Prediction of water quality time series data based on least squares support vector machine

This paper proposes using least squares support vector machine (LS-SVM) algorithm to construct a non-linear time series forecasting model for predicting water quality.

Proposed System

The Wastewater Monitoring project will consist of two modules. The first module will be responsible for monitoring the water quality by providing visualizations of physical, chemical and biological factors taken from the dataset. The second module will be responsible for predicting the water quality using machine learning, based on the current data.

In the following sections, the properties of the dataset are given and the two planned modules are explained.

Water Quality Monitoring System

The governmental agency, The Ministry of Environment and Urban Planning, needs a reporting system that visualizes the observations from important rivers, lakes, and marine special environmental protection areas (SEPA) in Turkey.

The reports are then used for decision making. In the past, the reports and visualizations in these reports were prepared manually. Our goal is to develop a web-based reporting system for SEPA that will automatically read the observation data set and present test results to decision-makers. This system will have detailed filtering features and will be able to perform data visualization. The data collected from the field will be used to produce visualizations for statistical modeling. Data visualization is helpful to decision-makers in identifying features that are not easily noticed by statistical models or humans, such as detection of outlier values of parameters, missing values. Visualizations enable correlation analysis, determination of the relationship between dependent variables.

The tables were often plotted with a bar chart. The graphics were created from samples taken from the inlets and outlets of wastewater treatment plants, as well as from different points of seas, lakes, and rivers. The charts parameters vary by year and region, but generally depend on the parameters of pH, Temperature (0C), Light Transmittance (m), Dissolved Oxygen (mg / L), O2 (%), Ammonia (mg / L), Total Phenol (mg / L), Total Coliform (CFU / 100mL), Fecal Coliform (CFU / 100mL), Fecal Streptococcus(CFU / 100mL), Oil-Grease (mg / L), Color (Pt-Co), Fragrance (TON).

There are some studies on water quality visualization. One of them used old-fashioned interactive maps and various types of plotting. Besides, tables are used, which contain something similar to the parameters used to determine water quality in the reports provided to us by the agency, sorted by years with parameter values of total phosphorus, total nitrogen, electrical conductivity, pH, dissolved oxygen.

A modern user-friendly interface, more effective and easy-to-understand graphics will be produced by considering the types of graphics and parameters used previously. Also, the locations where the test sample was taken will be displayed on the interactive maps.

Tools and Frameworks

There are multiple promising options for generating charts and creating user interfaces. The tools that are being considered include (but not limited to): Electron, React.js, Chart.js, ASP.NET, Flask, Qt and more. We are weighing all the options and will be considering the input of the agency since the end product will run on their environment.

Water Quality Prediction

The second goal of the project is to predict the future properties of a water sample. These predicted properties can then be used to predict the water quality and inform the water treatment plant.

For the problem of forecasting how to treat the water, since the quality of water can be affected by various parameters and such parameters show a complex non-linear relationship with each other and water quality, traditional techniques for data processing are no longer efficient enough.

Tools and Frameworks

For this project, we decided to use the Python programming language for implementing machine learning algorithms. The machine learning models will be trained locally. As for the machine learning framework, we decided to use

Tensorflow because, after training, the model can be easily used in Tensorflow.js, therefore making it very easy to run in a browser.

REFERENCES

[1] APEC. 'The History of Clean Drinking Water', 2018. [Online]. Available: https://www.freedrinkingwater.com/resource-history-of-clean-drinking-water.htm [Accessed: 2020/11/01]

[2] Minnesota Department of Health, 'Bacteria, Viruses, and Parasites in Drinking Water', 2019. [Online PDF]. Available: https://www.health.state.mn.us/communities/environment/water/docs/contaminants/parasitesfactsht.pdf [Accessed: 2020/11/01]

[3] A. N. Ahmed, F. B. Othman, H. A. Afan, R. K. Ibrahim, C. M. Fai, M. S. Hossain, M. Ehteram, and A. Elshafie, "Machine learning methods for better water quality prediction," Journal of Hydrology, vol. 578, p. 124084, Aug. 2019.

[4] J.-T. Kuo, M.-H. Hsieh, W.-S. Lung, and N. She, "Using artificial neural networks for reservoir eutrophication prediction," Ecological Modelling, vol. 200, no. 1-2, pp. 171–177, 2007. Retrieved from: https://www.sciencedirect.com/science/article/abs/pii/S0304380006002985?via%3Dihub

[5] A. Zaqoot, A. K. Ansari, M. A. Unar, and S. H. Khan, "Prediction of dissolved oxygen in the Mediterranean Sea along Gaza, Palestine – an artificial neural network approach," Water Science and Technology, vol. 60, no. 12, pp. 3051–3059, 2009. Retrieved from: https://iwaponline.com/wst/article-abstract/60/12/3051/13774/Prediction-of-dissolved-oxygen-in-the?redirectedFrom=fulltext

[6] Sengorur, B , Dogan, E , Koklu, R , Samandar, A . "Dissolved Oxygen Estimation using Artificial Neural Network for Water Quality Control", Electronic Letters on Science and Engineering 1 pp. 13-16, 2005. Retrieved from: https://dergipark.org.tr/en/pub/else/issue/29326/313793

[7] L. Xu and S. Liu, "Study of short-term water quality prediction model based on wavelet neural network," Mathematical and Computer Modelling, 22-Dec-2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895717712003676. [Accessed: 01-Nov-2020].

[8] Tan, G., Yan, J., Gao, C. and Yang, S. “Prediction of water quality time series data based on least squares support vector machine”, Procedia Engineering, 31, pp.1194-1199. 2012.
[9] Unwin, A. (2020). Why is Data Visualization Important? What is Important in Data Visualization? Harvard Data Science Review, 2(1). Retrieved from: https://doi.org/10.1162/99608f92.8ae4d525
[10] Ramsay, Ian & Shen, S. & Tennakoon, S.. (2009). Water Quality Visualisation and Tracking - Generic Decision Support Tool. Retrieved from: https://www.researchgate.net/publication/237627349_Water_Quality_Visualisation_and_Tracking_-_Generic_Decision_Support_Tool