A Comparison of Machine Learning Methods for Water Quality Prediction

Inshira Anwar, Karin Farajnejadi, Ethan Hawkins,Sephora Yameogo

Computer Science, University of Houston

Abstract

With the global decline in aquatic life populations and accessibility to clean water, water quality monitoring is imperative to protect the health of all aquatic life and the communities that depend on them globally. For example, in Texas, many waterways are exposed to contamination from diverse sources. To raise awareness on the matter, this study focussed on predicting the future trend of the most important determining factor of water quality: Dissolved Oxygen. To do so, the Long Short-Term Recurrent Neural Network(LSTM RNN) and the Support Vector Regression (SVR) were used on a time series dataset from the Water Quality Database provided by the United States Geological Survey (USGS). The models tested varied in the pre-processing methods used to construct them, notably scalers, Principal Component Analysis (PCA), and Kernel Principal Component Analysis (KPCA). The various impacts of these preprocessing methods were heavily examined by utilizing the following metrics: Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2). The results of this experiment show that across both the LSTM RNN models and the SVR models, PCA with standard scaling returned the best results across all metrics. Still, the best-performing model was the SVR using PCA with standard scaling which returned an RMSE of 0.268 and an MAE of 0.216, suggesting a superior fit for dissolved oxygen prediction on a USGS dataset.

Background

Although Machine Learning is commonly used in water quality monitoring, traditional methods still struggle to yield optimal results, especially on big data. To address this issue, Smail Dilmi and Mohamed Ladja proposed the use of deep learning algorithms such as the LSTM RNN, and feature extraction techniques such as LDA, PCA, and ICA to improve the accuracy of real-time water quality monitoring. The best models from their experience were the LSTM & LDA and LSTM & ICA. They both delivered an accuracy score of 99.72%.

Inspired by their work, this study experimented with diverse machine learning methods to predict the concentration of Dissolved Oxygen (DO) available to marine organisms in Texas's streams and rivers using a dataset from the United States Geological Survey Water.

Methods

For the experiment, 4 different preprocessing methods and 2 different learning algorithms were used to predict the amount of dissolved oxygen and compare the impact of those different methods on the quality of our predictions.

First, the data was scaled with the standard scaler and the Robust scaler to, respectively, fit all predictors in the same range and reduce the impact of outliers. Then, PCA and KPCA were performed to establish independence among our features. Resulting of this step was the four different training and testing sets used to implement the LSTM RNN and SVR model, thus having a total of 8 different models to compare. **(Figure 1**)

The Long Short Term Memory Recurrent Neural Network (LSTM RNN), was chosen for its ability to avoid the vanishing gradient problem, especially when working with a high-frequency time-series dataset. The Support Vector Regression(SVR) was chosen for its ability to perform non-linear regression. .

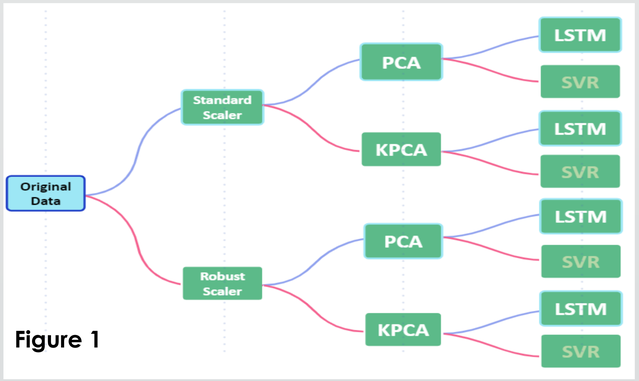

Figure 1

Future Direction

As Water quality continues to be a significant concern, the power of machine learning has proven to be a huge influence on environmental studies and prediction. Future research work to improve real-time water quality prediction includes and is not limited to:

(1) Testing different pre-processing methods

(2) Applying different LSTM or RNN models

(3) Utilizing different features

Results

The potential of the LSTM and SVR models was evaluated by measuring the evolution of the metrics and loss functions MSE, RMSE, MAE, and R-Squared, focusing on performance evaluation and model optimization. **Figure 2 & 3** provide visual aid of the fitting and predictions from the best LSTM and SVR models in orange, against the true observed values of dissolved oxygen.

For the evaluation metrics of the LSTM models, all numbers were very decent and reasonable outcomes, yielding low values, all close to the reference point of 0, which indicates a good fit. It can be observed that across all metrics, the LSTM using PCA & standardized scaling provided the smallest loss, with an RMSE 0.346.

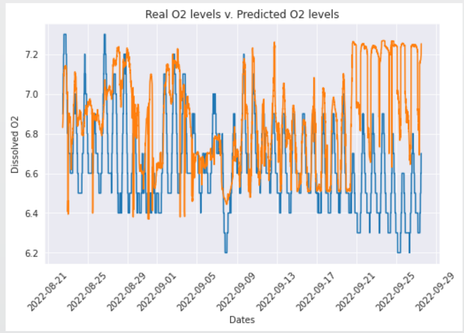

Figure 2: Observed values vs LSTM PCA SD predictions

Globally, the SVR models also delivered low metrics scores, with the SVR using PCA & standardized scaling outperforming the best LSTM model by yielding the lowest metrics values with an RMSE of 0.268. This model used a Gaussian kernel function, implying that the relationship between the predictors of Dissolved Oxygen is non-linear. This result is surprising considering the notoriety of the LSTM RNN and neural networks in general when it comes to time series data and water quality monitoring.

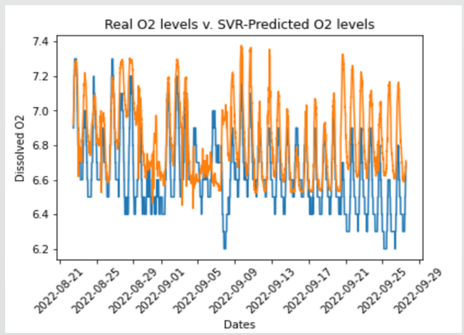

Figure 3: Observed values vs SVR PCA SD predictions

Conclusion

The goal of this study was to infer the impact of diverse machine learning and features extraction techniques on water quality monitoring by predicting the concentration of dissolved oxygen in rivers and streams. After evaluating eight different experimental procedures based on the LSTM RNN and SVR, it can be concluded that the best fit for our data was the Support Vector Regression with PCA & standard scaling. This model provided the lowest error rate with an RMSE of 0.268, indicating an adequate representation of the relationship between dissolved oxygen and the features of our dataset. Therefore, this model is fit to successfully generate future trends in dissolved oxygen.

Acknowledgments

Thanks to Dr. Nouhad Rizk for the opportunity to present our findings and the guidance provided throughout our work, the USGS for providing free reliable data, and authors S. Dilmi and M. Ladjal for the motivation of our work.