KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)
Terakreditasi Nomor 204/E/KPT/2022 | Vol. 5, No. 1, Januari (2024), pp. 182-188

Advancing River Water Quality Prediction A Comparative Study of Anomaly Detection Techniques for Optimizing Dissolved Oxygen Level Forecasting

Gregorius Airlangga
Information System Department, Universitas Katolik Indonesia Atma Jaya, Indonesia
E-mail: gregorius.airlangga@atmajaya.ac.id

Abstract

In the realm of environmental monitoring, particularly river water quality, the study at hand addresses the paramount challenge of accurately predicting dissolved oxygen (DO) levels—a critical indicator of aquatic ecosystem health. This research targets the complexities inherent in environmental datasets, including the presence of anomalies that can skew predictive models, thereby undermining the reliability of DO level forecasts. By applying and critically evaluating advanced anomaly detection methods—One-Class SVM, Isolation Forest, and Autoencoders—the study endeavors to enhance predictive accuracy and address gaps in existing research methodologies. The methodology encompasses data collection, preprocessing, anomaly detection, and evaluation, working with a dataset comprising five indicators across eight monitoring stations. The research process entailed thorough data preparation, ensuring dataset integrity and uniformity. Anomaly detection was meticulously performed, with each method revealing varying outlier detection sensitivities. The One-Class SVM method identified 23 outliers, the Isolation Forest found 38, and the Autoencoders flagged 88. When assessing the impact on model accuracy, reflected by the RMSE, the Isolation Forest method outperformed the others, achieving the lowest RMSE of 0.9668, indicating a more effective anomaly mitigation contributing to a cleaner dataset. In contrast, the Autoencoders, while detecting the most anomalies, yielded the highest RMSE, suggesting a propensity to overfit and misclassify data variations as anomalies. This study illuminates the criticality of selecting suitable anomaly detection methods tailored to the dataset's nuances, emphasizing that the choice profoundly influences predictive model performance. The Isolation Forest's proficiency in this context underscores its potential as a robust method for environmental data analysis, capable of balancing outlier detection accuracy with predictive model precision.

Keywords: *Environmental, Data Science, Isolation Forest, Anomaly Detection*

1. Introduction

Environmental monitoring, with a specific focus on river water quality, plays a crucial role in ecological conservation and public health [1]–[3]. Among various water quality parameters, dissolved oxygen (DO) levels stand out due to their direct impact on aquatic life and the overall health of river ecosystems [4]–[6]. Accurately predicting DO levels is imperative for effective environmental management. However, the complexity of environmental data, characterized by multi-variable interactions and anomalies, poses significant challenges to predictive accuracy [7]–[9]. The presence of anomalies in the dataset, whether due to natural variability, environmental incidents, or measurement errors, can significantly affect the reliability of predictive models [10]–[12]. A substantial body of research has been dedicated to environmental monitoring and water quality analysis. Traditional studies have often employed statistical and machine learning models to predict various water quality parameters, including DO levels [13]–[15]. These studies

ISSN: 2720-992X

have provided valuable insights into the factors influencing water quality and the potential applications of predictive models in environmental monitoring.

However, existing works reveal a notable gap in the handling of anomalies within environmental datasets. While some studies have acknowledged the challenge of outliers in environmental data, there has been limited exploration into systematic, comprehensive methods for anomaly detection and mitigation. Most existing approaches have either overlooked the complexity of these anomalies or applied standard outlier removal techniques without a thorough comparison of their effectiveness in the specific context of environmental data [16]–[18]. Furthermore, the impact of anomaly detection on feature selection and model performance has not been extensively studied in the realm of environmental science. This gap indicates a need for research that not only applies advanced anomaly detection methods to environmental datasets but also evaluates how these methods influence the overall predictive modeling process.

This study aims to bridge this gap by applying a state-of-the-art approach that compares multiple advanced anomaly detection techniques, including One-Class SVM, Isolation Forest, and Autoencoders. This approach is novel in its comprehensive application and comparison of these methods specifically in the context of river water quality datasets. By addressing the identified gaps in existing research, this study contributes to the field of environmental data science by enhancing the accuracy and reliability of predictive models for water quality parameters, particularly dissolved oxygen levels. This approach also sets a precedent for future research in the field, suggesting new directions and methodologies for handling complex environmental datasets. Building on this identified need, the study aims to:

a) Apply and compare various anomaly detection methods for their effectiveness in identifying and removing outliers from environmental datasets.
b) Assess the impact of anomaly removal on dataset quality and the relationships between different water quality indicators.
c) Re-evaluate the feature selection process post-anomaly detection to optimize the predictive model for DO levels.
d) Analyze the effect of these preprocessing steps on the performance of a RandomForestRegressor model, using metrics like Root Mean Squared Error (RMSE) to gauge improvements.

The paper is structured to provide a comprehensive overview of the methodology, followed by a detailed presentation of the results from the applied anomaly detection techniques. The analysis section will delve into the implications of these results for environmental monitoring and predictive modeling, culminating in a conclusion that synthesizes the key findings and offers directions for future research.

2. Reseach Methodology

In this work, we follow several steps for doing experiments. The steps include data collection, feature preprocessing, anomaly detection, and Evaluation. The dataset has five indicators that are measured at 8 stations of the state water monitoring system.
Indicators of river water quality in this dataset are:

a) $O2_{(i)}$: Dissolved oxygen (O2) is measured in mgO2/cub. dm (i.e. milligrams of oxygen (O2) in the cubic decimeter).
b) $NH4_{(i)}$: Ammonium ions (NH4) concentration is measured in mg/cub. dm (i.e. milligrams in the cubic decimeter).
c) $NO2_{(i)}$: Nitrite ions (NO2) concentration is measured in mg/cub. dm (i.e. milligrams in the cubic decimeter).
d) $NO3_{(i)}$: Nitrate ions (NO3) concentration is measured in mg/cub. dm (i.e. milligrams in the cubic decimeter).

e) $BOD5_{(i)}$: Biochemical oxygen demand, which is determined in 5 days ("BOD5" or "BOD"). BOD5 is measured in MgO/cub. dm (i.e. milligrams of oxygen in cubic decimeters).

2.1. Data Collection and Preparation

Three distinct datasets, 'sample_submission.csv', 'train.csv', and 'test.csv', are utilized. The 'train.csv' dataset includes training data vital for model development, 'test.csv' contains data for validation, and 'sample_submission.csv' serves as a format guide for outputs. Columns named "Id" in the original datasets are renamed to "id" to ensure consistency across all datasets, facilitating easier data manipulation and analysis. An extensive initial examination of the datasets is conducted to gain insights into the basic dataset structure, including data types, column names, and the presence of missing values. This step is crucial for planning subsequent data cleaning and processing strategies.

a) Initial Dataset Handling: Define the original training and testing datasets as D_{train} and D_{test}, respectively. Renaming operation: $D_{train['Id']} \rightarrow D_{train['id']}$ and $D_{test['Id']} \rightarrow D_{test['id']}$ to ensure uniform column names across datasets.
b) Computational Formulae for Missing Values: For each column c in D_{train} and the sample submission dataset, calculate the missing value percentage as follows: $m_{train}(c) = 100 \times (1 - (\text{count of non-null in c in } D_{train} / \text{total rows in } D_{train}))$ $m_{data}(c) = 100 \times (1 - (\text{count of non-null in c in the sample submission / total rows in the sample submission}))$

A detailed comparison is made between the missing values in the original datasets and the 'sample_submission.csv' dataset. This step is critical for assessing data integrity and planning appropriate data cleansing or imputation strategies. The 'train.csv' dataset undergoes a cleansing process to eliminate missing values, ensuring the reliability of the dataset for further analysis.

2.2. Feature Importance

The cleaned 'train.csv' dataset is merged with 'sample_submission.csv', creating a comprehensive dataset for subsequent feature importance analysis. Utilizing a RandomForestRegressor model, the study conducts a detailed analysis of feature importance. Parameters such as the number of estimators (1000) and maximum depth (7) are fine-tuned to optimize the model. The model is trained on the combined dataset, and the importance of each feature in predicting dissolved oxygen levels is calculated. The results of the feature importance analysis are visually presented through a horizontal bar chart. This visualization facilitates an intuitive understanding of which features have the most significant impact on the model's predictive capability.

a) Model Definition and Training: Define a RandomForestRegressor model M with parameters: $n_{estimators} = 1000$, $d_{max} = 7$. Train M on the combined dataset and extract feature importance for each feature fi in the feature set $F = \{f_1, f_2, ..., f_n\}$.
b) Mean Absolute Error (MAE) Calculation: The MAE is calculated to validate the model's predictions where $MAE = (1/N) \Sigma |y_i - \hat{y}_i|$ where y_i are the actual values, $\hat{y}_i$ are the predicted values, and N is the number of observations.

2.3. Anomaly Detection Techniques

a) One-Class SVM Anomaly Detection: Anomaly score s(x) for a data point x is defined as $s(x) = sgn(f(x) - \rho)$. This anomaly detection technique is applied to identify outliers in the dataset. The method operates by fitting a model that captures the normal data distribution, and outliers are identified as data points that deviate significantly from this distribution.
b) Isolation Forest for Outlier Detection: Average path length h(x) is utilized to detect anomalies, with shorter paths indicating outliers. Implemented with a

specific contamination parameter, this technique identifies anomalies based on their ease of isolation from the rest of the dataset. It's particularly effective for datasets with a mix of normal and abnormal data points.

c) Autoencoder-Based Anomaly Detection: Reconstruction error $e(x) = \|x - \hat{y}_x\|^2$ is used to identify outliers, where $\hat{y}_x$ is the reconstructed output. An autoencoder neural network is constructed to learn the normal data distribution. By training the network to reconstruct the input data and measuring the reconstruction error, anomalies are identified as data points with high reconstruction errors.

d) Outlier Marking and Removal: Each technique marks the detected outliers, which are then excluded from the dataset. This process results in a refined dataset, free from distortions caused by anomalous data points.

2.4. Evaluation

a) Iterative Feature Addition and Evaluation: The feature selection process is dynamic and iterative. Starting with the most crucial feature, additional features are gradually included based on their importance ranking. At each step, the model is re-evaluated to assess the impact of the added feature.

b) Model Performance Monitoring: Throughout the feature addition process, the performance of the RandomForestRegressor model is closely monitored using RMSE. This metric provides a quantitative measure of the model's prediction accuracy at each step.

c) Determination of Optimal Feature Set: The feature addition process continues until there is no significant improvement in RMSE, indicating the point at which the optimal set of features for the model has been identified.

d) Strategy for Unbiased Assessment: To validate the robustness and reliability of the model, performance evaluations are conducted using various random seeds. This approach ensures that the model's performance is not biased by a particular random state.

e) Comprehensive Performance Metrics Analysis: The RMSE is used as the primary metric to assess the model's performance. A detailed comparison of RMSE scores, both before and after the implementation of anomaly detection and feature selection, is carried out.

f) Efficacy and Improvement Assessment: This comparative analysis aims to quantify the improvements in model performance attributable to the anomaly detection and feature selection methodologies. The results of this analysis are crucial for demonstrating the effectiveness of the applied techniques in enhancing predictive accuracy.

g) Iterative Feature Addition and RMSE Evaluation: At each step i, use a feature subset Fi to train model M and compute RMSE where $RMSE = \sqrt{(1/N)\ \Sigma\ (y_i - \hat{y}_i)^2}$ Continue until RMSE improvement is minimal, determining the optimal feature set.

h) Robust and Unbiased Model Assessment: Compute RMSE multiple times with different random seeds. Report the average RMSE as the final performance metric, ensuring an unbiased and robust evaluation of the model's accuracy.

3. Results and Discussion

The table presents results from the application of three distinct anomaly detection techniques: Support Vector Machine (SVM), Isolation Forest, and Autoencoders, on a dataset aimed at predicting dissolved oxygen levels. The SVM, a method based on the concept of decision planes that define decision boundaries, identified 23 outliers, with a resulting RMSE of 1.2073. This relatively higher RMSE suggests that while SVM was

conservative in marking outliers, it may not have captured all the true anomalies, possibly due to its sensitivity to the choice of kernel and hyperparameter settings, which could have led to a suboptimal separation of outliers from normal data points.

In comparison, the Isolation Forest method detected a larger number of outliers, 38, and yielded a notably lower RMSE of 0.9668. Isolation Forest is known for its effectiveness in isolating anomalies rather than profiling normal data points, which can be particularly advantageous for datasets with numerous and diverse anomalies. Its performance, as evidenced by the lowest RMSE, indicates a robust detection capability that likely contributed to a cleaner dataset, leading to improved predictive accuracy.

Autoencoders, utilizing neural networks to reconstruct data, identified the highest number of outliers at 88 but resulted in the highest RMSE of 1.2881. This outcome suggests that the Autoencoders were perhaps overly sensitive, flagging too many points as outliers. This could be due to the Autoencoders' threshold settings for reconstruction error, which, if not adequately tuned, can misclassify normal variations in data as anomalies. Consequently, this may have led to the removal of valuable information, adversely affecting the model's ability to make accurate predictions.

Each method's efficacy is inherently dependent on the underlying distribution and nature of the dataset, including the presence and type of anomalies. The Isolation Forest's superior performance in this scenario could be attributed to its non-parametric approach, which does not assume an underlying distribution for normal data points and is typically more flexible in accommodating the dataset's unique characteristics. The results collectively highlight the importance of selecting the right anomaly detection technique based on the dataset's specific attributes and the desired balance between identifying true outliers and retaining predictive accuracy.

Table 1. Comparison Result

Metric/Method	SVM	Isolation Forest	Autoencoders
Number of Detected Outliers	23	38	88
RMSE	1.2073	0.9668	1.2881

4. Conclusion

This study embarked on a comparative analysis of three anomaly detection techniques—Support Vector Machine (SVM), Isolation Forest, and Autoencoders—to enhance the predictive accuracy of a model aimed at estimating dissolved oxygen levels in river systems. The findings illustrate a marked variation in the performance of these techniques, as evidenced by the number of detected outliers and the associated Root Mean Squared Error (RMSE) values. The SVM technique demonstrated a conservative approach, identifying the fewest outliers. Its performance, resulting in a moderate RMSE, suggests that it may be suitable for datasets where maintaining the integrity of data is crucial, and the cost of false positives — wrongly identified outliers — is high. However, the SVM's effectiveness can be significantly influenced by the selection of hyperparameters and kernel choice, which necessitates careful tuning to optimize model accuracy.

The Isolation Forest emerged as the most proficient technique in this study, detecting more outliers than SVM and improving the model's RMSE to the lowest value among the three methods. This success underscores the strength of the Isolation Forest in dealing with complex, multi-dimensional datasets and its capability to enhance model performance without extensive parameter tuning. Its non-parametric nature offers a flexible approach that is less susceptible to the assumptions of data distribution, making it robust in identifying true anomalies. Conversely, the Autoencoders, despite identifying

the highest number of outliers, resulted in the least favorable RMSE. This indicates a potential over-sensitivity to deviations in the data, leading to an overfitting scenario where the model may have discarded valuable information, mistaking it for anomalies. While Autoencoders are powerful in learning intricate data patterns, their performance is heavily contingent on the appropriate calibration of the network architecture and error thresholding, which requires a deep understanding of the dataset's characteristics.

In conclusion, the Isolation Forest method proved to be the most effective in this context, striking a balance between outlier detection and predictive accuracy. The study's findings advocate for a methodical approach to selecting anomaly detection techniques, tailored to the specificities of the dataset in question. Future work could expand upon these results by exploring hybrid models or ensemble techniques that combine the strengths of individual methods to further refine outlier detection and improve predictive accuracy. This research contributes to the field of environmental data science by providing insights into the application and efficacy of different anomaly detection methods. It reinforces the significance of method selection in the preprocessing phase, which is pivotal for the development of accurate and reliable predictive models in ecological monitoring and assessment.

References

[1] S. Giri, “Water quality prospective in Twenty First Century: Status of water quality in major river basins, contemporary strategies and impediments: A review,” *Environ. Pollut.*, vol. 271, p. 116332, 2021.

[2] A. Pal, Y. He, M. Jekel, M. Reinhard, and K. Y.-H. Gin, “Emerging contaminants of public health significance as water quality indicator compounds in the urban water cycle,” *Environ. Int.*, vol. 71, pp. 46–62, 2014.

[3] M. A. Sadat, Y. Guan, D. Zhang, G. Shao, X. Cheng, and Y. Yang, “The associations between river health and water resources management lead to the assessment of river state,” *Ecol. Indic.*, vol. 109, p. 105814, 2020.

[4] A. Csábrági *et al.*, “Estimation of dissolved oxygen in riverine ecosystems: Comparison of differently optimized neural networks,” *Ecol. Eng.*, vol. 138, pp. 298–309, 2019.

[5] S. Gheorghe *et al.*, “Metals toxic effects in aquatic ecosystems: modulators of water quality,” *Water Qual.*, vol. 87, pp. 59–89, 2017.

[6] J. Bir, M. S. Sumon, and S. M. B. Rahaman, “The effects of different water quality parameters on zooplankton distribution in major river systems of Sundarbans Mangrove,” *IOSR J. Environ. Sci. Toxicol. Food Technol.*, vol. 11, pp. 56–63, 2015.

[7] A. N. Matheri, F. Ntuli, J. C. Ngila, T. Seodigeng, and C. Zvinowanda, “Performance prediction of trace metals and cod in wastewater treatment using artificial neural network,” *Comput. \& Chem. Eng.*, vol. 149, p. 107308, 2021.

[8] Y. Bai and J. Zhao, “A novel transformer-based multi-variable multi-step prediction method for chemical process fault prognosis,” *Process Saf. Environ. Prot.*, vol. 169, pp. 937–947, 2023.

[9] S. Dragović, “Artificial neural network modeling in environmental radioactivity studies--A review,” *Sci. Total Environ.*, vol. 847, p. 157526, 2022.

[10] A. A. Cook, G. M\is\irl\i, and Z. Fan, “Anomaly detection for IoT time-series data: A survey,” *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6481–6494, 2019.

[11] R. Kromanis and P. Kripakaran, “SHM of bridges: characterising thermal response and detecting anomaly events using a temperature-based measurement interpretation approach,” *J. Civ. Struct. Heal. Monit.*, vol. 6, pp. 237–254, 2016.

[12] S. F. Gould *et al.*, “A tool for simulating and communicating uncertainty when modelling species distributions under future climates,” *Ecol. Evol.*, vol. 4, no. 24,

pp. 4798–4811, 2014.
[13] K. Chen *et al.*, "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data," *Water Res.*, vol. 171, p. 115454, 2020.
[14] A. N. Ahmed *et al.*, "Machine learning methods for better water quality prediction," *J. Hydrol.*, vol. 578, p. 124084, 2019.
[15] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, M. Maashi, and others, "Water quality prediction using artificial intelligence algorithms," *Appl. Bionics Biomech.*, vol. 2020, 2020.
[16] S. Thudumu, P. Branch, J. Jin, and J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *J. Big Data*, vol. 7, pp. 1–30, 2020.
[17] A. Blázquez-Garc\'\ia, A. Conde, U. Mori, and J. A. Lozano, "A review on outlier/anomaly detection in time series data," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–33, 2021.
[18] M. N. K. Sikder and F. A. Batarseh, "Outlier detection using AI: a survey," *AI Assur.*, pp. 231–291, 2023.