2024 Scienceline Publication

Journal of Civil Engineering and Urbanism

Volume 14, Issue 3S: 278-282; September 15, 2024 ISSN-2252-0430

DOI: https://dx.doi.org/10.54203/jceu.2024.29

Enhancing Water Level Prediction through a Hybrid Feature Selection Approach

Thalosang Tshireletso[1,4]✉, Yashon Ouma[1], Ditiro Moalafhi[3], and George Anderson[2]

[1]*Department of Civil Engineering, University of Botswana, Gaborone, Botswana*

[2]*Department of Computer Sciences, University of Botswana, Gaborone, Botswana*

[3]*Department of Wildlife and Aquatic Resources, Botswana University of Agriculture and Natural Resources, Sebele, Botswana*

[4]*Department of Civil Engineering, University of Cape Town, Cape Town, South Africa*

✉Corresponding author's Email: tshireletsot@ub.ac.bw

RESEARCH ARTICLE

PII: S225204302400029-14

Received: June 25, 2024

Revised: September 02, 2024

Accepted: September 05, 2024

ABSTRACT

Accurate prediction of water levels (WL) is essential for various applications, from flood management to environmental monitoring. In this study, an enhanced approach to feature selection tailored for water level prediction models is presented. Our method integrates Mutual Information and Recursive Feature Elimination with Cross-Validation (RFECV), augmented by the Non-Dominated Sorting Genetic Algorithm II (NSGA-II), to systematically evaluate and refine subsets of features. Mutual Information facilitates the identification of relevant feature dependencies, while RFECV iteratively eliminates less informative features to optimize predictive accuracy. The inclusion of NSGA-II further enhances the selection process by considering multiple conflicting objectives simultaneously, such as maximizing R^2 score and minimizing the number of selected features, RMSE, and MAE. Through extensive experimentation and validation on real-world datasets, we demonstrate the effectiveness of our hybrid feature selection approach in capturing intricate relationships within the data, leading to significantly improved predictive performance in water level prediction models.

Keywords: Mutual Information, Recursive Feature Elimination with Cross validation, Non-Dominated Sorting Genetic Algorithm II

INTRODUCTION

In the era of big data and complex datasets, the task of accurately predicting outcomes or trends from data has become increasingly vital across various domains (Li et al., 2019; Saeys et al., 2007). Regression modelling stands as a cornerstone in numerous fields, from finance to healthcare, where understanding and forecasting numerical outcomes are paramount (Guyon and Elisseeff, 2003). However, amidst the abundance of available features, selecting the most relevant ones that contribute significantly to the predictive power of a model becomes crucial. This necessity forms the crux of our study, as we delve into the realm of feature selection and its implications on regression modelling accuracy.

The significance of feature selection reverberates throughout the landscape of machine learning, impacting not only the performance of models but also their interpretability and computational efficiency (Li et al., 2019; Dash and Liu, 1997). By distilling datasets to their most informative attributes, feature selection mitigates the curse of dimensionality, alleviating issues such as overfitting and enhancing model generalization. Moreover, in domains where resource constraints or interpretability are paramount, selecting a parsimonious set of features aids in building more comprehensible and deployable models.

The objective of this paper is to explore the efficacy of feature selection techniques in enhancing the performance of regression models. To achieve this objective, we adopt a two-fold approach: firstly, we investigate traditional feature selection methods such as Mutual Information and Recursive Feature Elimination with Cross-Validation (RFECV), evaluating their impact on model accuracy and feature subset size. Subsequently, we employ multi-objective optimization techniques, leveraging the Non-dominated Sorting Genetic Algorithm II (NSGA-II), to identify Pareto-optimal solutions that balance model accuracy and feature subset complexity (Deb, et al., 2002; Kohavi and John, 1997).

Through this endeavour, we aim to provide insights into the trade-offs inherent in feature selection and

To cite this paper: Tshireletso T, Ouma Y, Moalafhi D, and Anderson G (2024). Enhancing Water Level Prediction through a Hybrid Feature Selection Approach. *J. Civil Eng. Urban.*, 14 (3S): 278-282. DOI: https://dx.doi.org/10.54203/jceu.2024.29

empower practitioners with knowledge to make informed decisions when building regression models. By elucidating the interplay between feature selection methods and model performance, this study contributes to advancing the understanding of best practices in regression modelling, thereby facilitating more robust and interpretable predictive analytics solutions.

Background and related work

The field of feature selection and regression modelling is a vast and intricate domain, and numerous studies have been carried out, scrutinizing various techniques and their effectiveness across a wide array of domains. Prominent among these are the comparative studies carried out by Kohavi and John (1997). They evaluated the performance of wrapper methods, particularly recursive feature elimination (RFE), with a keen focus on enhancing the accuracy of regression models. Their findings were enlightening, showing significant improvements in R-squared values and a decrease in mean squared error when RFE was applied in contrast to the baseline models.

In the same vein, Dash and Liu in 1997 embarked on a deep exploration of the application of filter methods, particularly mutual information-based feature selection. They showcased how these techniques bring about notable enhancements in predictive accuracy and model interpretability across a myriad of regression tasks. Their work served as a seminal contribution to the field, and it has been widely referenced in subsequent studies.

In recent times, hybrid approaches have gained considerable traction with the aim of synergizing the strengths of different feature selection techniques. An exemplary work in this regard is that of Hsu et al. in 2011. They proposed a ground breaking framework that seamlessly integrates filter and wrapper methods. The performance of their hybrid approach was superior in terms of both accuracy and computational efficiency. It achieved state-of-the-art results on benchmark datasets, surpassing the performance of individual feature selection methods.

In addition to these academic advancements, domain-specific applications have also showcased the practical utility of feature selection. An excellent example of this is in the finance sector, where researchers have utilized feature selection to identify key predictors for stock price forecasting models. This has led to more accurate predictions and has informed investment decisions (Yang, et al., 2019), contributing significantly to the field of financial forecasting.

Healthcare is another sector where feature selection techniques have played a pivotal role. They have been instrumental in identifying biomarkers and clinical predictors for disease diagnosis and prognosis (Saeys et al., 2007). By selecting relevant features from high-dimensional medical datasets, researchers have achieved remarkable accuracies in predicting patient outcomes and guiding personalized treatment strategies.

Furthermore, in environmental science, feature selection has been applied to remote sensing data for land cover classification and ecological modelling (Heman et al., 2013). These applications underscore the versatility and efficacy of feature selection in addressing real-world challenges across various domains.

Through empirical validation and rigorous evaluation, these studies have not only advanced our understanding of feature selection techniques but also provided practical insights into their application and impact on regression model performance. By leveraging the collective knowledge generated from these works, our study aims to contribute further to the evolving landscape of feature selection and regression modelling, ultimately enhancing predictive accuracy and interpretability in data-driven decision-making processes.

MATERIALS AND METHODS

Experimental Setup

The MI-RFECV-NSGA-II feature selection method to the test using various predictor variables and a target variable. To ensure we had a clean dataset, we handled missing values and encoded categorical variables. We split the dataset into training and testing sets, using the 80-20 standard for training and testing respectively, and kept a consistent random seed for reproducibility.

A combination of Mutual Information (MI), Recursive Feature Elimination with Cross-Validation (RFECV), and the Non-Dominated Sorting Genetic Algorithm II (NSGA-II) for feature selection. MI and RFECV helped us identify a subset of relevant features. The NSGA-II algorithm was then used to optimize the feature subset by maximizing the coefficient of determination (R^2) and minimizing the selected features. Afterwards, Random Forest (RF) regression model was trained using the features identified by the NSGA-II algorithm.

The RF model was trained and tested using different performance metrics, such as R^2 score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), which were used to judge its accuracy and generalization

capabilities. Finally, we compared how our RF model, trained with features selected by NSGA-II, stood up against RF models trained with other feature selection techniques like MI and RFECV. This comparison gave us a better understanding of how effective the MI-RFECV-NSGA-II feature selection method is in pinpointing informative features for regression tasks.

Mutual information (MI)

Mutual information measures the dependency between two variables, in this case, each feature and the target variable 'WL' (water level). The mutual information between a feature Xi and the target variable Y is calculated as:

$$I(X_i;Y) = \sum_{x_i \epsilon X_i} \sum_{y \epsilon Y} p(x_i, y) log\left(\frac{p(x_i,y}{p(x_i)\cdot p(y)}\right) \quad (1)$$

Where:

- $p(x_i, y)$ is the joint probability mass function of X_i and Y
- $p(x_i)$ and $p(y)$ are the marginal probability mass functions of X_i and Y respectively.

Features with higher mutual information scores are considered more informative.

Recursive feature elimination with cross-validation (RFECV)

The RFECV recursively removes less informative features and selects the subset that optimizes model performance, typically measured using cross-validation. The process involves training the model with the current set of features and evaluating its performance. Features with the lowest importance, often determined by their coefficients in the model, are pruned iteratively until the desired number of features is reached. The selection process can be represented as:

$$arg_{features}^{min}(MSE_{CV}) \quad (2)$$

Where:

- MSE_{CV} is the mean squared error evaluated through cross-validation.

Multi-objective optimization with NSGA-II

NSGA-II aims to optimize four conflicting objectives simultaneously: maximizing R^2 (coefficient of determination), minimizing the number of selected features, minimizing RMSE, and minimizing MAE. Let's denote these objectives as f1, f2, f3, and f4 respectively.

The output of NSGA-II consists of a set of Pareto-optimal solutions, denoted as Pareto, representing trade-offs between R^2, the number of selected features, RMSE, and MAE. Each solution in Pareto represents a unique combination of these objectives, providing insights into the optimal feature subsets that balance predictive accuracy, model simplicity, and error metrics.

RESULTS AND DISCUSSION

The initial Random Forest (RF) model performed well, but improved after applying feature selection techniques, table 1. The RF model with Mutual Information Recursive Feature Elimination with Cross-Validation (MI-RFECV) showed increased R^2 and decreased RMSE, despite a slight increase in MAE. Further enhancement was achieved with the addition of the NSGA-II optimization algorithm, leading to significant performance improvements. This underlines the effectiveness of feature selection, especially MI-RFECV-NSGA-II, in improving RF model accuracy and reducing errors.

Table 1 - Results for RF without feature selection, RF with MI-RFECV and RF with RFECV-NSGA-II

Model	R^2	RMSE	MAE
RF without Feature selection	0.735	9.680	5.314
RF with MI-RFECV	0.804	8.311	6.769
RF with MI-RFEV-NSGA-II	0.896	4.485	2.598

Table 2 - Comparison of the study with recent studies

Method	R^2	RMSE	MAE
RF with MI-RFECV-NSGA	0.896	4.485	2.598
Chamlal et al (2023)	0.870	-	-
Sandru & David (2019)	-	8.930	6.250
El Touati et al (2023)	-	5.270	4.910
Hsu et al (2011)	0.860	7.540	-
Al-Aghbari et al (2022)	-	5.680	-
Sun et al (2021)	-	4.810	3.230

The study results highlight the effectiveness of feature selection techniques in improving the predictive accuracy of regression models, especially for environmental data analysis. A significant improvement in model performance when using Mutual Information Recursive Feature Elimination with Cross-Validation (MI-RFECV) was seen. This improvement was further amplified when integrating the Non-Dominated Sorting Genetic Algorithm II (NSGA-II) in the feature selection process.

Recent studies, Table 2, align with the findings of this research, emphasizing the role of advanced feature

selection and optimization in achieving higher predictive accuracy.

Chamlal et al. (2023) proposed a two-step feature selection procedure for handling high-dimensional data, focusing on balancing computational efficiency and prediction performance. While their method demonstrates substantial improvement in model accuracy, it does not address the simultaneous optimization of conflicting objectives, such as minimizing RMSE and MAE, which our MI-RFECV-NSGA-II approach successfully achieves.

Șandru and David (2019) introduced a unified feature selection and hyperparameter optimization framework based on Bayesian methods. Although their approach streamlines regression modelling, the lack of integration with multi-objective optimization frameworks limits its ability to balance prediction performance against feature subset complexity, as achieved in our study. In comparison, MI-RFECV-NSGA-II provides a comprehensive solution, enhancing generalization and predictive performance.

El Touati et al. (2023) presented an adaptive feature selection method that dynamically refines feature subsets in machine learning models. While effective in improving computational efficiency and accuracy, their approach lacks the ability to optimize multiple objectives simultaneously. By contrast, our method explicitly incorporates multi-objective optimization through NSGA-II, achieving an R^2 of 0.896 while significantly reducing RMSE and MAE.

In the context of environmental and hydrological modelling, Al-Aghbari et al. (2022) demonstrated a hybrid multi-objective optimization approach for water flooding applications. Their work highlights the potential of multi-objective algorithms in environmental systems but focuses more on physical processes than feature selection.

Similarly, Sun et al. (2021) explored adaptive surrogate modelling for hybrid optimization, showcasing the advantages of multi-objective optimization in constrained scenarios. These studies reinforce the importance of incorporating multi-objective frameworks, which our MI-RFECV-NSGA-II method leverages for improved regression performance.

Overall, this study demonstrates that the integration of feature selection techniques like MI-RFECV with advanced multi-objective optimization algorithms such as NSGA-II significantly outperforms traditional and recent hybrid methods. This is evident in the remarkable improvement in R^2, RMSE, and MAE metrics, highlighting the robustness and applicability of the proposed approach for complex predictive modelling tasks.

CONCLUSION

The study was conducted on feature selection in regression modelling, with a special focus on environmental data analysis. After careful experimentation and evaluation, the following conclusions were reached about the effectiveness of feature selection techniques and their influence on model performance.

It's clear that feature selection is key to improving the predictive accuracy of regression models. Traditional methods like Mutual Information and Recursive Feature Elimination with Cross-Validation (MI-RFECV) are quite good at identifying relevant features and boosting performance metrics such as the R^2 score, RMSE, and MAE. However, when the Non-Dominated Sorting Genetic Algorithm II (NSGA-II) was incorporated into the feature selection process, the MI-RFECV-NSGA-II approach led to significant improvements in predictive accuracy. It delivered higher R^2 scores, and reduced error rates compared to the standard methods.

Comparing the results archived to results from previous studies, it's evident that MI-RFECV-NSGA-II outperforms traditional feature selection methods. While earlier methods have shown positive results in improving model accuracy, this study takes it a step further. Leveraging the optimization capabilities of NSGA-II to identify more optimal feature subsets. The resulting enhancement in model performance metrics suggests that MI-RFECV-NSGA-II provides a more structured and efficient approach to feature selection, especially in high-dimensional datasets and complex regression tasks.

DECLARATIONS

Corresponding author

Correspondence and requests for materials should be addressed to Thalosang Tshireletso; E-mail: tshireletsot@ub.ac.bw; ORCID: 0000-0002-5112-8077

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Acknowledgements

This research project was funded by both the USAID Partnerships for Enhanced Engagement in Research

(PEER) under the PEER program cooperative agreement number: AID-OAA-A-11- 00012 and the University of Botswana, Office of Research and Development (ORD).

Authors' contribution

Conceptualization and Methodology; Thalosang Tshireletso; Supervision; Yashon Ouma, Ditiro Moalafhi and Geroge Anderson. Funding Acquisition; Yashon Ouma. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no conflict of interest.

REFERENCES

Ahmed, I., Abu, M., & El-Henawy, I. (2017). A Feature Selection Algorithm based on Mutual Information using Local Non-uniformity Correction Estimator. *International Journal of Advanced Computer Science and Applications*, 8(6). DOI:10.14569/IJACSA.2017.080656

Al-Aghbari, M., Gujarathi, A., Al-Wadhahi, M., & Chakraborti, N. (2022). Hybrid Multi-objective Optimization Approach in Water Flooding. *Journal of Energy Resources Technology,145*(3),032103. https://doi.org/10.1115/1.4052623

Chamlal, H., Benzmane, A., & Ouaderhman, T. (2023). A Two-Step Feature Selection Procedure to Handle High-Dimensional Data in Regression Problems. *2023 International Conference on Decision Aid Sciences and Applications(DASA)*. https://doi.org/10.1109/DASA59624.2023.10286637

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis, 1*(1–4), 131–156. https://doi.org/10.1016/S1088-467X(97)00008-5

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002) - A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182-197. DOI:10.1109/4235.996017

El Touati, Y., Ben Slimane, J., & Saidani, T. (2023). Adaptive Method for Feature Selection in the Machine Learning Context. *Engineering, Technology & Applied Science Research,13*(2),123–130. https://doi.org/10.48084/etasr.7401

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182. Available at http://jmlr.csail.mit.edu/papers/volume3/guyon03a/guyon03a.pdf.

Herman, G., Zhang, B., Wang, Y., & Ye, G. (2013). Mutual information-based method for selecting informative feature sets. *Pattern Recognition*, 46(12), 3315-3327. DOI:10.1016/j.patcog.2013.04.021

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2), 273-324. DOI: 10.1016/S0004-3702(97)00043-X.

Hsu, H.-H., Hsieh, C.-W., & Lu, M.-D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications*, 38(7), 8144-8150. DOI: 10.1016/j.eswa.2010.12.156.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517. DOI: 10.1093/bioinformatics/btm344.

Şandru, E.-D., & David, E. (2019). Unified Feature Selection and Hyperparameter Bayesian Optimization for Machine Learning-Based Regression. *International Symposium on Signals, Circuits and Systems (ISSCS)*. https://doi.org/10.1109/ISSCS.2019.8801728

Sun, R., Duan, Q., & Mao, X. (2021). A Multi-Objective Adaptive Surrogate Modelling-Based Optimization Algorithm for Constrained Hybrid Problems. *Environmental Modelling & Software, 144*, 105272. https://doi.org/10.1016/j.envsoft.2021.105272