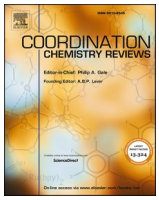
Coordination Chemistry Reviews 534 (2025) 216583

Contents lists available at ScienceDirect

Coordination Chemistry Reviews

journal homepage: www.elsevier.com/locate/ccr



Enhancing machine learning in gas–solid interaction analysis: Addressing feature selection and dimensionality challenges

ARTICLE INFO

Keywords:
Machine learning
Gas-solid interactions
Feature selection
Bias
Dimensionality reduction
Statistical methods

ABSTRACT

This paper addresses the critical importance of accurate analysis in research, emphasizing the necessity of error-free and unbiased calculations. While ground truth values are pivotal for validating accuracy, their absence poses challenges in feature importance, feature selection, and clustering methods commonly used in machine learning. Liu et al. have introduced innovative models targeting gas-solid interactions, but their reliance on model-specific methodologies raises concerns about potential biases and erroneous conclusions. This study advocates for robust statistical validation techniques, including the application of Variance Inflation Factor (VIF), Spearman's correlation, and Kendall's tau, to enhance the reliability of feature selection and ensure more accurate insights. By emphasizing a rigorous approach to statistical significance, this paper aims to improve the interpretability and effectiveness of machine learning applications in this specialized field.

Accurate analysis relies on researchers performing calculations that are both error-free and unbiased. Recognizing the presence of ground truth values is essential for validating accuracy. In supervised machine learning, ground truth values are typically accessible, facilitating straightforward accuracy validation. However, methods like feature importance, feature selection, and clustering frequently do not have clear ground truth references, which complicates the assessment of accuracy. As a result, it is vital for researchers to rigorously validate statistical significance to ensure reliable results. This paper demonstrates the application of feature selection within analytical processes and offers insights into its usefulness, despite the challenges posed by the absence of ground truth values in certain situations.

Liu et al. introduced innovative machine learning models focused on gas-solid interaction materials and devices, contributing to advancements in this specialized field. The machine learning (ML) workflow typically comprises four essential stages: dataset preparation, feature selection, model training, and model evaluation [1]. Within this framework, feature selection aims to maximize the correlation between features and predicted properties, minimize inter-feature correlations, and ensure ease of feature acquisition. Commonly employed methods for feature selection include Filter, Wrapper, and Embedded approaches. In addition to extracting inherent features, the creation of new features is vital for continuously optimizing ML models. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), play a crucial role in transforming high-dimensional data into a more manageable low-dimensional space. By simplifying data structures, PCA highlights underlying distributions and relationships, employing orthogonal transformations to convert linearly related features into a limited number of independent variables that retain most of the essential information from the original dataset [1].

While Liu et al. have made commendable advancements in developing machine learning models for gas-solid interactions, this paper raises critical concerns regarding their approach to feature selection and the correlation between features and predicted properties. The model-specific nature of their methodologies may lead to erroneous conclusions if not subjected to rigorous validation. It is essential for Liu et al. to acknowledge the significance of ground truth values in validating their outcomes. While supervised machine learning techniques benefit from these values for accuracy assessment, the fields of feature importance and feature selection often lack such references, complicating the evaluation of their effectiveness. The absence of ground truth can result in varied methodologies for calculating feature importance across different models, ultimately introducing biases. In fact, over 100 peer-reviewed articles have documented non-negligible biases in feature importance derived from various models [2–6].

Furthermore, while Liu et al. have proposed Principal Component Analysis (PCA) for feature reduction, it is vital to recognize that PCA's linear and parametric nature may inadequately address the complexities found in non-linear and nonparametric data [7,8]. Given the absence of ground truth values in feature importance calculations, three key components warrant careful consideration: the distribution of data, the statistical relationships between variables, and the validation of these relationships through statistical significance tests using *p*-values. To address these challenges, this paper advocates for employing robust statistical methods that are capable of accommodating non-linearity and nonparametric characteristics, such as Spearman's correlation [9] and Kendall's tau [10], both supplemented with *p*-values for rigorous validation. By integrating these methodologies, Liu et al. could enhance the reliability and interpretability of their ML models, leading to more accurate insights into gas-solid interactions.

This paper recognizes that machine learning (ML) serves as a valuable tool for target predictions, yet it critically examines the reliance on biased feature importances generated by ML models. To enhance the integrity of feature selection, the application of Variance Inflation Factor (VIF) is essential for identifying and removing features exhibiting collinearity and interactions [11]. By addressing these correlations

https://doi.org/10.1016/j.ccr.2025.216583
Received 28 January 2025; Accepted 23 February 2025

proactively, VIF helps to mitigate feature inflation and ensures a more accurate representation of the underlying relationships. Subsequently, this paves the way for implementing robust statistical methods that can reveal true associations between the target variable and its features. By emphasizing the importance of rigorous feature evaluation and reduction, this paper advocates for a more nuanced and reliable approach to leveraging ML models in predictive analysis.

Funding

This research has no fund.

Authors' contribution

Yoshiyasu Takefuji completed this research and wrote this article.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Code availability

Not applicable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] P.Y. Liu, X. Zhu, X. Ran, H. Bi, X. Huang, N. Gu, Machine learning for gas–solid interaction materials and devices, Coord. Chem. Rev. 524 (2025) 216329, https://doi.org/10.1016/j.ccr.2024.216329.
[2] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: learning a Variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (2019) 177.
[3] A. Demircioğlu, Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics, Insights Im. 12 (1) (2021) 172. Published 2021 Nov 24, https://doi.org/10.1186/s13244-021-01115-1.
[4] J. Krawczuk, T. Łukaszuk, The feature selection bias problem in relation to high-dimensional gene data, Artif. Intell. Med. 66 (2016) 63–71, https://doi.org/10.1016/j.artmed.2015.11.001.
[5] C. Strobl, A.L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinformatics. 8 (2007) 25, https://doi.org/10.1186/1471-2105-8-25.
[6] T. Salles, L. Rocha, M. Gonçalves, A bias-variance analysis of state-of-the-art random forest text classifiers, ADAC 15 (2021) 379–405, https://doi.org/10.1007/s11634-020-00409-4.
[7] M. Chen, K. Papadikis, C. Jun, N. Macdonald, Linear, nonlinear, parametric, and nonparametric regression models for nonstationary flood frequency analysis, J. Hydrol. 616 (2023) 128772, https://doi.org/10.1016/j.jhydrol.2022.128772.
[8] S.W. Jarantow, E.D. Pisors, M.L. Chiu, Introduction to the use of linear and nonlinear regression analysis in quantitative biological assays, Current Protocols 3 (6) (2023) e801, https://doi.org/10.1002/cpz1.801.
[9] H. Yu, A.D. Hutson, A robust spearman correlation coefficient permutation test, Commun. Stat. Theory Methods 53 (6) (2024) 2141–2153, https://doi.org/10.1080/03610926.2022.2121144.
[10] K. Okoye, S. Hosseini, Correlation tests in R: Pearson Cor, Kendall's tau, and Spearman's rho, in: R Programming, Springer, Singapore, 2024, https://doi.org/10.1007/978-981-97-3385-9_12.
[11] J. Jacob, R. Varadharajan, Robust variance inflation factor: a promising approach for collinearity diagnostics in the presence of outliers, Sankhya B 86 (2024) 845–871, https://doi.org/10.1007/s13571-024-00342-y.

Yoshiyasu Takefuji
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan
E-mail address: takefuji@keio.jp.