Contents lists available at ScienceDirect

Journal of Energy Chemistry

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

http://www.journals.elsevier.com/journal-of-energy-chemistry/

Letter

Unveiling hidden biases in machine learning feature importance

Yoshiyasu Takefuji

Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan

ARTICLE INFO

Article history:
Received 6 October 2024
Revised 21 October 2024
Accepted 21 October 2024
Available online 5 November 2024

Keywords:
Machine learning
Feature importance
Potential bias
Chi-squared and P-value

Nirmal et al. presented a machine learning-based design of ternary organic solar cells, utilizing feature importance [1]. This paper highlights the alarming potential biases in the use of feature importance in machine learning, which can lead to incorrect conclusions and outcomes. Many scientists and researchers including Nirmal et al. are unaware that feature importances in machine learning in general are model-specific and do not necessarily represent true associations between the target and features.

While machine learning aims to accurately predict the target using features, feature importances are merely a byproduct of the process, and their values can vary significantly across different models. In other words, different models generate different feature importances, even though true associations between the target and features exist independently of the models used. Nirmal et al. employed five machine learning models: classification and regression tree (CART), random forest (RF), gradient boosting (GB), linear models (LMs), and artificial neural networks (ANNs). For instance, a feature deemed highly significant by RF may receive a lower ranking in LM, thereby obscuring true underlying relationships despite existing associations.

Nirmal et al. demonstrated that individual machine learning models, such as CART, RF, GB, LM, and ANN, exhibited distinct sets of feature importances. These variations suggest biases rather than true associations.

This paper identifies biases of feature importance in machine learning. Feature importance in machine learning can have potential biases due to several reasons. Firstly, feature importance values can vary significantly across different models. For instance, tree-based models like Random Forests and Gradient Boosting can give different importance scores compared to linear models like Logistic Regression. This variability can lead to inconsistent interpretations of which features are truly important [2].

Additionally, features with more levels or categories can appear more important in tree-based models because they have more opportunities to split the data and reduce impurity, which can give an inflated sense of their importance and lead to biased interpretations [3]. The training data used to build the model can also introduce biases; if the training data is not representative of the real-world scenario, the feature importance derived from it can be misleading [4]. Some algorithms might inherently favor certain types of features over others. For example, algorithms that rely heavily on correlation might overemphasize features that are highly correlated with the target variable, even if they are not causally related [5]. Lastly, complex models, such as deep learning models, often lack transparency, making it difficult to understand the true importance of features, which can lead to biased interpretations and decisions based on the model's output [6]. SHAP relies on a machine learning model, which means that the feature importances and explanations it provides are inherently influenced by the model's specific characteristics. Using a different model will result in different explanations from SHAP. In other words, feature importances derived from machine learning models tend to be biased. To obtain true associations, feature importances should be calculated independently of any machine learning models.

Instead of biased feature importance in machine learning, Chi-squared tests and *P*-values play a crucial role in calculating true

E-mail address: takefuji@keio.jp

https://doi.org/10.1016/j.jechem.2024.10.032
2095-4956/

associations between the target and features, model independent. Chi-Squared Tests are a collection of non-parametric statistical methods used to examine whether there is a significant association between categorical variables. These tests assess how well the observed data fit with the expected data under the assumption of no association. There are several types of Chi-Squared Tests, including the Chi-Squared Test of Independence, the Chi-Squared Goodness-of-Fit Test, and the Chi-Squared Test for Homogeneity [7].

The Chi-Squared Test of Independence is utilized to determine whether two categorical variables are independent of each other. On the other hand, the Chi-Squared Goodness-of-Fit Test evaluates whether the observed frequency distribution of a single categorical variable matches an expected distribution. The Chi-Squared Test for Homogeneity is similar to the test of independence but is used to compare the distribution of a categorical variable across different populations, such as comparing product preferences among various age groups.

The procedure for performing a Chi-Squared Test begins with formulating hypotheses. The Null Hypothesis (H_0) typically states that there is no association between the variables (they are independent), while the Alternative Hypothesis (H_1) suggests that there is an association. The Chi-Squared statistic (χ^2) is then calculated using the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where (O_{ij}) is the observed frequency and (E_{ij}) is the expected frequency.

The degrees of freedom, which for a test of independence is calculated as (number of rows -1) $\times$ (number of columns -1), are determined next. The calculated χ^2 value is compared to a critical value from the Chi-Squared distribution table based on the degrees of freedom. If $\chi 2$ exceeds the critical value, the null hypothesis is rejected, indicating a significant association between the variables. It's important to note that Chi-Squared Tests assume that the data are in frequencies or counts, categories are mutually exclusive, and the sample size is sufficiently large, typically with expected frequencies of at least 5.

Moving on to *P*-Values, a *P*-Value is a probability measure that quantifies the evidence against a null hypothesis. Specifically, it represents the likelihood of obtaining results as extreme as those observed, assuming that the null hypothesis is true. In hypothesis testing, a low *P*-Value (typically $\leq$ 0.05) suggests that the observed data are unlikely under the null hypothesis, leading to its rejection in favor of the alternative hypothesis. Conversely, a high *P*-Value(> 0.05) indicates insufficient evidence to reject the null hypothesis, suggesting that the observed data are consistent with it [8,9]. There is no strict rule for setting a predetermined significance level; it is often determined by the specific context and conditions of the study. In many cases, a *P*-value threshold of less than 0.05 or 0.01 is commonly used to indicate strong evidence against the null hypothesis. These levels are selected based on conventional standards in the field, the research question, the consequences of Type I and Type II errors, and the specific needs of the analysis.

The process of utilizing a *P*-Value involves several steps. First, hypotheses are set up, with the null hypothesis (H_0) stating no effect or difference, and the alternative hypothesis (H_1) proposing that an effect or difference exists. A significance level (α), commonly set at 0.05 or 0.01, is chosen to determine the threshold for rejecting H_0. A test statistic relevant to the specific test being used is then computed. The *P*-Value is calculated based on this test statistic, representing the probability of observing the data—or something more extreme—under the null hypothesis. If the *P*-Value is less than or equal to α, the null hypothesis is rejected; if it is greater, the null hypothesis is not rejected.

Chi-Squared Tests and *P*-Values are closely related in statistical analysis. When performing a Chi-Squared Test, the calculated Chi-Squared statistic is used to determine the corresponding *P*-Value based on the Chi-Squared distribution with the appropriate degrees of freedom. This *P*-Value is then used to decide whether to reject the null hypothesis of independence or goodness-of-fit. For instance, in a Chi-Squared Test of Independence, after calculating the χ^2 statistic and degrees of freedom, the *P*-Value is found using the Chi-Squared distribution. If the *P*-Value is less than or equal to the chosen significance level α, the null hypothesis is rejected, indicating an association between the categorical variables.

Understanding Chi-Squared Tests and *P*-Values is fundamental for making informed decisions in statistical hypothesis testing, particularly when working with categorical data. These concepts allow researchers to determine the likelihood that observed patterns are due to chance or represent a meaningful association or fit within the data.

Machine learning focuses on accurately predicting the target, while feature importances highlight the associations between the target and the features. The proposed statistical approach complements the machine learning methods without disrupting them. Prepare data and call Chi-squared function: chi2_contingency(data). Both Chi-squared tests and *P*-values are integral components of inferential statistics, aiding researchers in validating their hypotheses and assessing the strength of their findings.

The findings from Nirmal et al. and other similar studies highlight the need for caution when interpreting feature importance in machine learning models. The variability in feature importance values across different models can lead to inconsistent and potentially misleading conclusions about which features are truly important. This has significant implications for the design and optimization of ternary organic solar cells, as well as other applications in the field of energy chemistry. Researchers and scientists must be aware that feature importances are model-specific and do not necessarily represent true associations between the target and features. This awareness can help in making more informed decisions and avoiding incorrect conclusions that could impact the development and performance of solar cells and other technologies.

This paper does not aim to discourage the use of machine learning. Instead, it seeks to provide important cautions to researchers regarding the potential biases associated with feature importance in machine learning models. By highlighting these concerns, the paper encourages researchers to adopt a more critical and informed approach when interpreting feature importance, ensuring that their conclusions and outcomes are based on robust and reliable analyses. Ultimately, this awareness can lead to more accurate and meaningful applications of machine learning across various fields.

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.

References

[1] K.A. Nirmal, T.D. Dongale, S.S. Sutar, A.C. Khot, T.G. Kim, J Energy Chem. 100 (2025) 337–347.
[2] M. Openja, G. Laberge, F. Khomh, Empir Software Eng 29 (2024) 22.
[3] M. Saarela, S. Jauhiainen, SN Appl. Sci. 3 (2021) 272.
[4] D. Theng, K.K. Bhoyar, Knowl. Inf. Syst. 66 (2024) 1575–1637.

[5] Henriques J, Rocha T, de Carvalho P, Silva C, Paredes S. Interpretability and Explainability of Machine Learning Models: Achievements and Challenges. In: Pino E, Magjarević R, de Carvalho P, eds. International Conference on Biomedical and Health Informatics 2022. ICBHI 2022. IFMBE Proceedings. Vol 108. Cham: Springer; (2024).

[6] M. Frasca, D. La Torre, G. Pravettoni, et al., Discov. Artif. Intell. 4 (2024) 15.

[7] Z. Chen, Sci. Rep. 12 (2022) 3158.

[8] D. Berrar, Data Min. Knowl. Disc. 36 (2022) 1102–1139.

[9] I.H. Sarker, SN Comput. Sci. 2 (2021) 160.