GOOD SCIENCE
BETTER MEDICINE
BEST PRACTICE

ANNALS OF ONCOLOGY
DRIVING INNOVATION IN ONCOLOGY

LETTER TO THE EDITOR

Chi-square and *P*-values versus machine learning feature selection

Fraunhoffer et al.[1] used the least absolute shrinkage and selection operator (LASSO) and random forest (RF) methods for feature selection, which may not be ideal. They noted that incorporating master regulator transcripts from the neoplastic cell phenotype played a significant role in the LASSO feature selection for all drugs, with the highest proportions from gemcitabine and 5-fluorouracil.[1] Feature selection in machine learning may not, however, provide true associations.[2-4] Instead, chi-square tests and *P*-values should be used to ensure true associations, rather than relying on LASSO and RF methods.[5-7] Consequently, their results may differ.

Feature selection in machine learning may not provide true associations for several reasons. One major issue is overfitting, where models, especially complex ones, capture noise rather than true underlying patterns in the training data. Additionally, machine learning algorithms often identify correlations between features and the target variable, but these correlations may not imply causation. This distinction is crucial because a correlation does not necessarily mean that one variable causes the other.

Another challenge is the bias and variance inherent in feature selection methods. These methods can be sensitive to the specific data used, leading to biased or high-variance results that do not generalize well to new data. Furthermore, different algorithms have different strengths and weaknesses. For example, LASSO may shrink some coefficients to zero, potentially missing important features, while RF may overemphasize certain features due to their inherent structure.

Chi-square tests and *P*-values, however, are statistical methods that provide true associations between the target and features. Chi-square tests and *P*-values measure the statistical significance of the association between features and the target variable, helping to distinguish true associations from random noise. These methods are grounded in hypothesis testing, providing a framework to test whether the observed associations are likely to be due to chance. Additionally, statistical methods can control for confounding variables, ensuring that the associations identified are not spurious. Finally, results from statistical tests are generally reproducible and can be validated across different datasets.

When using chi-square tests and *P*-values for feature selection, a higher chi-square value signifies a stronger association between the feature and the target variable.

Y. Takefuji*

Faculty of Data Science, Musashino University, Koto-ku, Tokyo, Japan
(*E-mail: takefuji@keio.jp).

Available online xxx

https://doi.org/10.1016/j.annonc.2024.10.013
DOI of original article: https://doi.org/10.1016/j.annonc.2024.06.010

FUNDING

None declared.

DISCLOSURE

The author has declared no conflicts of interest.

REFERENCES

1. Fraunhoffer N, Hammel P, Conroy T, et al. Development and validation of AI-assisted transcriptomic signatures to personalize adjuvant chemotherapy in patients with pancreatic ductal adenocarcinoma. *Ann Oncol*. 2024;35(9):780-791.
2. Theng D, Bhoyar KK. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowl Inf Syst*. 2024;66:1575-1637.
3. Dhal P, Azad C. A comprehensive survey on feature selection in the various fields of machine learning. *Appl Intell*. 2022;52:4543-4581.
4. Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci*. 2021;2:160.
5. Jariyavajee C, Lamjiak T, Ratanasanya S, et al. Cash stock strategies during regular and COVID-19 periods for bank branches by deep learning. *PLoS One*. 2022;17(6):e0268753.
6. Spencer R, Thabtah F, Abdelhamid N, Thompson M. Exploring feature selection and classification methods for predicting heart disease. *Digit Health*. 2020;6:2055207620914777.
7. Hossen MJ, Ramanathan TT, Al Mamun A. An ensemble feature selection approach-based machine learning classifiers for prediction of COVID-19 disease. *Int J Telemed Appl*. 2024;2024:8188904.