Report

The article titled ‘How doppelgänger effects in biomedical data confound machine learning’ focuses on the current state of doppelgänger effects in machine learning work in biomedicine. The following issues are mainly introduced. The article describes i. what doppelgänger effects are, ii. the abundance of doppelgänger effects in the field of biomedical informatics machine learning, iii. examples of analysis based on renal cell carcinoma (RCC) proteomics data, iv. existing and recommended methods to deal with doppelgänger effects. After reading through the article and thinking about it, I have a certain cognitive understanding of doppelgänger effects and some preliminary understanding of doppelgänger effects in biomedical machine learning. In the following, I will answer the questions given in the context with my own understanding

Why is the doppelgänger effects unique to biomedical data? First, I believe that the specificity of doppelgänger effects is reflected in its prevalence in biomedical data. doppelgänger effects have been observed in several biomedical fields, such as modern bioinformatics (detailed evaluation of existing chromatin interaction prediction systems)[1], established fields of bioinformatics (protein function prediction)[2] and drug discovery (QSAR)[3]. Secondly, The uniqueness of doppelgänger effects in biomedicine is reflected in the serious consequences of the lack of attention to them in the field of machine learning. Even though doppelgänger effects are present in a significant number of cases, their treatment is still not a standardized process. A significant number of researchers still do not identify and treat doppelgänger effects in biomedical data at the data processing stage, which leads to errors in the validation results and over-training effects falsely. Thirdly, doppelgänger effects is a problem that are not easy to resolve analytically and fully solved. Although previous studies have proposed some methods to deal with it, such as ordination methods, dupChecker [4] and pairwise Pearson's correlation coefficient (PPCC) [5], and some suggestions to deal with it, it still requires a comprehensive and rigorous strategy to deal with it for different research contexts, otherwise it will seriously affect the effectiveness of machine learning models. To summarize, doppelgänger effects are a unique, thorny problem in biomedical machine learning, which leads to performance inflation of machine learning models and thus prevents them from being properly evaluated and trained.

Regarding how to avoid doppelgänger effects in biomedical data, I think there are two perspectives. The first perspective is a technical analysis and processing based on the original data set, The first perspective is the statistical tests or technical analysis which is processed based on the original dataset, such as the PPCC method mentioned in the text, robust independent validation checks and performing data stratification. These methods are effective in most cases and can mitigate or eliminate doppelgänger effects, but on small data sets with a wide distribution of doppelgänger data, this method has limitations that make the information obtained by the model inadequate.

The second perspective is optimized the acquisition method for future data by analyzing the characteristics of doppelgänger data in the original data, such as sensitive features and similar data points. Therefore, a more complete range of data acquisition and more refined data features can be developed to avoid the emergence of statistically similar data.

In addition, I found a previous case of a doppelgänger effect study. On genome-wide analyses of cancer specimens, researchers often share or reuse specimens in follow-up studies. If duplicate expression profiles in public databases go unnoticed, they will have an impact on reanalysis. [6] In a quantitative study using the PPCC approach, researchers discovered that in genomic analysis of different cancer specimens, patients with cancer types with high paired PCC, such as thyroid cancer, have very similar expression profiles and are difficult to distinguish based on expression data alone. In contrast, substantial genomic mutations provide unique expression patterns that facilitate doppelgänger identification in cancer types with low PCC, such as bladder cancer.

Reference

[1] F. Cao, M.J. Fullwood, Inflflated performance measures in enhancer–promoter interaction-prediction methods, Nat Genet 51 (2019) 1196–1198.

[2] M.N. Wass, M.J. Sternberg, ConFunc: functional annotation in the twilight zone, Bioinformatics 24 (2008) 798–806.

[3] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, R.K. Tekade, Artifificial intelligence in drug discovery and development, Drug Discov Today 26 (2021) 80–93.

[4] Q. Sheng, Y. Shyr, X. Chen, DupChecker: a bioconductor package for checking high throughput genomic data redundancy in meta analysis, BMC Bioinform 15 (2014) 323.

[5] L. Waldron, M. Riester, M. Ramos, G. Parmigiani, M. Birrer, The Doppelgänger effect: hidden duplicates in databases of transcriptome profifiles, J Natl Cancer Inst 108 (2016) djw146.

[6] Levi Waldron, Markus Riester, Marcel Ramos, Giovanni Parmigiani, Michael Birrer, The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles, *JNCI: Journal of the National Cancer Institute*, Volume 108, Issue 11, November 2016, djw146, https://doi.org/10.1093/jnci/djw146