Question 3 : doppelgänger effects

Introduction

Machine learning (ML) methods have been increasingly used in biomedical fields. When assessing the performance of a classifier in ML, the training and test datasets should be independently derived. However, the reliability of such methods could still be affected by the presence of data doppelgängers, which occur when independently derived data are very similar to each other, causing models to perform well regardless of how they are trained.

Data doppelgängers are abundant in modern bio-informatics. Given the potential of doppelgänger effects to confound machine learning, it is crucial to be able to identify the presence of data doppelgängers between training and validation sets before validation. Approaches like ordination methods or embedding methods are unfeasible because data doppelgängers are not necessarily distinguishable in reduced-dimensional space, while dupChecker method does not detect true data doppelgängers that are independently derived samples that are similar by chance. The prime limitation of another measure, the pairwise Pearson's correlation coefficient (PPCC), is that it never conclusively made a link between PPCC data doppelgängers and their ability to confound ML tasks. [1]

The confounding effects of PPCC data doppelgängers can be analyzed from a quantitative angle. The extent of this inflationary effect varies depending on two

main factors: the similarity of functional doppelgängers and the proportion of functional doppelgängers in the validation set. The more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance. [1]This points toward a dosage-based relationship between the number of PPCC data doppelgängers and the magnitude of the doppelgänger effect.

Examples and Analysis

1.Tumor purity estimation

Tumor purity estimation is crucial for accurate pathologic evaluation and for sample selection to minimize normal cell contamination in high throughput genomic analysis. A group of researchers [2] used data from ten different TCGA cohorts and a local Singapore cohort. Such international collaborations are beneficial to the cancer research community, but pose challenges to investigators developing independent validations and meta-analyses.

Re-use of tissue specimens is widespread in clinical genomic studies, creating doppelganger effect in publicly available datasets: hidden duplicates that, if left undetected, can inflate statistical significance or apparent accuracy of genomic models when combining data from different studies. [3]As genomic databases grow and collect tumor specimens from international collaborators, the chance of duplication increases. Analysis of duplicate samples is a substantial concern that could alter the identification of subsets of patients with clinical differences or the

development of specific gene signatures.

2.Path analysis

A study develops a methodological framework that integrates machine learning with path analysis to quantify behavioral pathways in bicycle-motor vehicle crashes.[4] The study used a dataset containing 9,296 bicycle-motor vehicle crashes to demonstrate the application of the framework. The path analysis chains machine learning models to establish the indirect linkages between contributing factors and injury severities through correlates of pre-crash behaviors. The framework is expected to support agencies' decision-making to improve cycling safety by reducing unsafe behaviors on roads.

However, this is likely to cause doppelgänger effects if the representativeness of the dataset used is not excellent enough, e.g., if the feature distribution of the dataset used for training is roughly within similar time periods or locations, the features of the validation and training sets will not differ much.

Way to avoid

1. Identify data doppelgängers before the training-validation split. The key aspects to identifying duplicates in a pair of datasets are:[3]

1) using transcript identifiers available in both datasets,

2) batch correction,

3) calculating Pearson’ s Correlation Coefficient (PCC) between every sample in one data set against every sample in the other dataset, and

4) duplicate-oriented outlier detection.

2.Perform careful cross-checks using meta-data as a guide. With this information from the meta-data, we are able to identify potential doppelgängers and assort them all into either training or validation sets, effectively preventing doppelgänger effects, and allowing a relatively more objective evaluation of ML performance. Or we can look for subsets of a validation set that are predicted correctly regardless of the ML method used.

3.Perform data stratification. Instead of evaluating model performance on whole test data, we can stratify data into strata of different similarities.

4.perform extremely robust independent validation checks involving as many data sets as possible, which ensures the diversity of datasets. Try to choose data from several categories to ensure that the data are not extremely similar, resulting in better model generalization.

5.Take a thorough review on potential collaborations and/or multiple institution clinical management of individual patients, as well as suspiciously similar clinical patient data and identifiers.

Conclusion

In this report, I show my understanding of doppelgänger effects, including causes of doppelgängers effects, limitations of methods to identify the presence of data doppelgängers and understanding on doppelgängers effects from a quantitative angle. I discuss in detail the doppelgänger effects in today's biomedical data through the case of Tumor purity estimation. Besides biomedicine, I also analyze data doppelgänger in path analysis and conclude that doppelganger effects are not unique to biomedical data. After integrating literature resource, I propose several methods to avoid doppelgänger effects.

References

[1] Wang L R, Wong L, Goh W W B. How doppelgänger effects in biomedical data confound machine learning[J]. Drug discovery today, 2021.

[2] Oner M U, Chen J, Revkov E, et al. Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study[J]. bioRxiv, 2021.

[3] Waldron, Levi; Riester, Markus; Ramos, Marcel; Parmigiani, Giovanni; Birrer, Michael. The doppelganger effect : Hidden duplicates in databases of transcriptome profiles[J]. Journal of the National Cancer Institute, Vol. 108, No. 11, djw146, 2016.

[4]Weike Lu, Jun Liu, Xing Fu, Jidong Yang, Steven Jones, Integrating machine learning into path analysis for quantifying behavioral pathways in bicycle-motor vehicle crashes[J], Accident Analysis & Prevention, Volume 168, 2022.