Q3

Introduction

Doppelgänger effects (DEs) in machine learning (ML) model is a situation used in evaluating the performance of machine learning model despite the training methods and quality. DEs in context of ML model first described by Wang et al. [1] is defined as a phenomenon where the training and testing sets measurement results are falsely high in similarity due to random faults, whereas the actual performance of the validation set may not be as successful. This can be seen in modern bioinformatics, which not only in ML models, but across the different biological data such as transcriptome profiles.

Owing to the potential risk of DEs leading to ineffective and unaware toxic drug screening discovery, DEs must be identified. There are currently two differentiate types, data doppelgängers (DDs) and functional doppelgängers (FDs). [2] Yet there is also complex in distinguishing between the actual well performance data and the DE-misleading data.

In this short review report, it will discuss on whether DEs are unique to biomedical data and if there are any methods to avoid DEs in the practice and development of health and medical science related ML models.

Is doppelgänger effect unique to biomedical data?

Doppelgänger effects initiates from the overlapping similarities between reference and validation data sets. Theoretically, it is high likely that other quantitative data sets may also encounter DEs. However, to this date, the DEs with the same context defined by Wang et al. [1] have yet to be recorded from quantitative research field other than biomedical data science. This may be due to that the impact caused by the DEs can be neglected but drug screening candidates need extra surveillance and management for the performance and results. Furthermore, as more sequencing and screening procedures are moved from manual to automatic in recent decades, DEs is still a relatively new concept encountered. Therefore, there are now relatively rising attention on this phenomenon especially in terms of biomedical data. It can be said that doppelgänger effect concept is unique to biomedical data in terms of the current literature recordings, yet it may not be unique in terms of the non-recognized occurrence of DEs in other fields.

DEs in other biomedical data types

Prior to the mentioning of DEs in ML models by Wang et al. [1], the same DEs have been mentioned by Waldron et al. [3] in the context of whole-genome analysis and transcriptome profiling in 2016. Waldron et al. conveys that the regeneration of public-available data samples in global clinical genomic studies can cause the inflating statistical significance or increase the seeming accuracy of genomic models. Since the intra-study and interstudy duplication occurrences will increase as more of the same database are being collected unknowingly.

DEs in quantitative angle

The principle factors of identifying DEs according to Waldron et al. [3] includes using transcript identifiers available in both datasets, batch adjustment and correction, pairwise Pearson's correlation coefficient, and duplicate-oriented outlier detection.

Pairwise Pearson's correlation coefficient (PPCC) compares all samples in the first dataset against all sample of the second data set, which describes the relationship between the sample pairs of different data sets. The higher the PPCC value means that data doppelgänger is observed from the pair. Yet only limited to identification of whether or not there is a doppelgänger sample. [1,3]

How to avoid DEs

To avoid DEs, potential DDs and FDs need to first be removed.[1] Yet purely removing the doppelgängers from the data sets is insufficient to prevent doppelgänger effects occurrence. Therefore, a following doppelgänger effects identification and verification need to be done. Wang and her team [2,4] in their most recent paper developed a PPCC DD identification and DD inflationary effect verification, which can be done using R programming. This allow the identification and verification within and between microarray and RNA-Sequencing datasets. In which, the PPCC DDs are then being proved as FDs. The reference value of the PPCC DDs cases must be extremely low or even better for 0. Other than using the PPCC model. The upregulation of quality controls in each steps of the experiments also need to be done.

Conclusion

Doppelgänger effects in the context of duplication similarity between reference and validation data sets is a relatively new concept in the ML models and the biomedical data field. Yet this doesn't necessarily mean that the phenomenon is unique only in the biomedical data field. Nonetheless, potential factors leading to DEs must be reduced and avoided to allow more effectives sequencing and screening. This can be done by removing the obvious duplications and with the PPCC identification as reference further identify the DDs and FDs to understand how they affect the overall statistical significant of the train-validation data results.

Reference

1. Wang LR, Wong L, Goh WWB. How doppelgänger effects in biomedical data confound machine learning. Drug Discovery Today. 2022; 27(3): p. 678-685. https://doi.org/10.1016/j.drudis.2021.10.017.

2. Wang LR, Choy XY, Goh WWB. Doppelgänger spotting in biomedical gene expression data. iScience. 2022; 25(8): p. 104788. https://doi.org/10.1016/j.isci.2022.104788.
3. Waldron L, Riester M, Ramos M, Parmigiani G, Birrer M. The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. J Natl Cancer Inst. 2016;108(11):djw146. Published 2016 Jul 5. http://doi.org/10.1093/jnci/djw146.
4. Wang LR, Fan XY, Goh WWB. Protocol to identify functional doppelgänges and verify biomedical gene expression data using doppelgangerIdentifier. STAR Protocols. 2022; 3(4): p. 101783. https://10.1016/j.xpro.2022.101783.