How doppelgänger effects in biomedical data confound machine learning

Drug Discovery Today, 2021

Highlights of this publication

Data doppelgängers occurs when independently derived training and validation sets are highly similar to each other, causing a machine learning model to perform well regardless of the quality of training (doppelgänger effect). Data doppelgängers are prevalent in biomedical dataset and they should be identified before model validation (Wang et al., 2021).

Introduction

Machine learning (ML) algorithms have increasingly been used as a diagnostic and prognostic tool in health and medical science. Despite the widespread use of ML models, clinical ML applications are often affected by confounders, defined as any variable causing spurious associations between the features and response variable. These confounders are often clinically irrelevant but can artificially increase the predictive performance of the algorithms. There are several methods to mitigate the presence of confounders (ie. variables such as age and gender), such as (i) matching samples in order to improve the balance of the data, (ii) using inverse probability weighting in neuro-imaging to weight the training samples in order to make the model better tailored to the population of interest (Linn et al., 2016; Rao et al., 2017), as well as (iii) using restricted permutations to detect, quantify and correct confounding in ML predictions (Neto, 2018).

Moreover, in the fields of biomedical research, the use of ML models has also accelerated the pace for drug discovery, leading to shorter drug development processes (Fleming, 2018). Given the expensive drug-testing process, it is thus of upmost importance that these ML models are correctly trained and tested to identify new drug candidates.

Despite the use of cross-validation techniques to properly evaluate these models, there exists other confounders such as data doppelgängers that could affect the reliability of such validation methods (Wang et al., 2021). Data doppelgängers affect the interpretation of the prediction outcome because the model will perform well regardless of their training quality. This observed doppelgänger effect is particularly unique in biomedical dataset, possibly because experimental or clinical results are usually performed in replicates and from multiple samples, and especially when the re-use of tissue specimens is widespread in clinical genomic studies, thus generating the doppelgänger effect in publicly available database. Therefore, the key to a more accurate machine learning predictive model relies on the ability of researchers to identify potential data doppelgängers beforehand.

Research in Context

In this publication, the authors first demonstrated the prevalence of data doppelgängers by using a benchmark renal cell carcinoma (RCC) with carefully designed controls. The RCC datasets is chosen due to its clear-cut negative cases (where doppelgängers are not permissible – sample pairs with different class labels), valid cases (where doppelgängers are permissible – sample pairs with

same class labels but from different samples) as well as positive cases (not doppelgängers because of data leakage - sample pairs with same class labels and same samples). Using the measure, Pairwise Pearson's Correlation Coefficient (PPCC), as a way to reveal the presence of doppelgängers (an anomalously high PPCC values denote a pair of data doppelgängers sample), the authors demonstrated that at least 50% of the samples are PCC data doppelgängers with at least one other sample (Wang et al., 2021).

Once the data doppelgängers are identified, the authors went on to determine if these doppelgängers are functional or not (ie. whether their presence have any confounding effect on the ML model). The authors found that the presence of data doppelgängers in both training and validation sets inflates the ML performance in a dosage-relationship (ie. the more the PPCC data doppelgängers, the higher the magnitude of the doppelgänger effect), across different datasets and ML models such as K-Nearest Neighbours, Naïve Bayes, Decision Tree as well as Logistic Regression Models. Moreover, the extent of doppelgänger effect depends on the similarity of functional doppelgängers and the proportion of functional doppelgängers in the validation set.

The authors also showed a way to mitigate these doppelgänger effect, by placing them all into either the training set or validation set, where the accuracy dropped to levels of expected accuracy of a model trained on random signatures, although this method is still a suboptimal solution. Other methods discussed include (i) using the PPCC outlier detection package (doppelgangR) to identify and remove the data doppelgängers in whole-genome analysis of cancer specimens to minimize their effects (Waldron et al., 2016), as well as to use (ii) several cell types to generate the training-evaluation pair for a more objective evaluation of ML performance (Cao & Fullwood, 2019). Despite all these mentioned methods, the complexity of data doppelgängers meant that even after removing all the highly correlated variables, the doppelgänger effect is still not resolved as the high correlations between sample pairs cannot be explained by a subset of highly correlated variables (Wang et al., 2021).

A list of recommendations is thus put forward by the authors to suggest ways to reduce the doppelgänger effects. The first is to use the meta-data as a guide to identify potential data doppelgängers, and to include them all into the training or validation sets, so that a more objective evaluation of the ML model performance can be obtained. Next is to stratify the datasets into several strata of different similarities (ie. PPCC data doppelgängers and non-PPCC data doppelgängers) can be used to evaluate the ML performance on each stratum separately. Third is to perform divergent validation so the it can inform about the objectivity of the classifier.

All in all, identifying and resolving data doppelgängers is an ongoing endeavor and in order to reduce ML performance inflation, one needs to check for potential data doppelgängers before assortment into training and validation sets.

REFERENCES

Cao, F., & Fullwood, M. J. (2019). Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nature Genetics*, *51*, 1196–1198.

Fleming, N. (2018). How artificial intelligence is changing drug discovery. *Nature*, *557*, S55–S57.

Linn, K. A., Gaonkar, B., Doshi, J., Davatzikos, C., & Shinohara, R. T. (2016). Addressing Confounding in Predictive Models with an Application to Neuroimaging. *The International Journal of Biostatistics*, *12*, 31–44.

Neto, E. C. (2018). Using permutations to detect, quantify and correct for confounding in machine learning predictions. 1–23. http://arxiv.org/abs/1805.07465

Rao, A., Monteiro, J. M., Mourao-Miranda, J., & Alzheimer's Disease Initiative. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, *150*, 23–49.

Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016). The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. *Journal of the National Cancer Institute*, *108*, 2–5.

Wang, L. R., Wong, L., & Goh, W. W. Bin. (2021). How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*.