Doppelgänger Effects on Machine Learning Models

Xinru SHAO

1 Abstract

Machine learning models are increasingly being used in drug development to enable faster identification of potential targets. This article focuses on examining doppelgänger effects found in machine learning models. It summarizes the causes and effects of doppelgänger effects, how to distinguish and reduce the consequences of data doppelgängers, and discusses whether doppelgänger effects only exist in the biomedical field. In addition, an example about cancer that includes doppelgänger effects is provided at the end of the article.

2 Introduction

In the current drug development process, machine learning is more and more widely used in the drug development process to screen candidate drug targets and other stages to greatly improve the efficiency of drug development and testing. In drug development, cross-validation is often used to evaluate mathematical models built in machine learning, and data doppelgängers in cross-validation methods usually affect the validation results. In turn, model evaluation is an essential part of the process of building qualified mathematical models. Data doppelgängers refers to the problem that the independent data in the data collection process are similar to each other, which in turn causes the test in the model validation process to show good detection results no matter what. In detail, in the process of evaluating a machine learning model, it is often necessary to divide the training set and the test set of the data set and test and compare the results separately to judge whether the model is successfully established, and the existence of data doppelgängers will cause the training set to It is highly similar to the test set, resulting in good results regardless of whether the model is trained well or not.

3 Phenomenon of data doppelgängers

3.1 How doppelgänger effects are caused

Today, with the development of modern bioinformatics, data doppelgängers have been observed in many cases: the prediction of chromosomal interaction mentioned in the literature, the prediction of protein function, etc., have a high degree of similarity between the training set and the test set. Due to the existence of data doppelgängers, it is often wrong conclusions that the model prediction results are highly accurate during the model testing process. Earlier, we stated that data doppelgängers are when samples appear similar across their measurements. However, functional doppelgängers are different from actual doppelgänger effects. Thus, we would like to understand better the level of similarity between suspected functional doppelgängers and the acceptable proportion of functional doppelgängers in the validation set.

3.2 How effective are data doppelgängers in confusing

The RCC experiments identified data doppelgängers in the PPCC data, which in turn determined to have a significant inflation effect on ML performance[1]. In the training and validation of the model, even if the features are randomly selected, the training process of PPCC shows a significant improvement in ML performance. In addition, the more doppelgänger pairs in the training and validation sets in the experiment, the more inflated the ML performance is, which fully demonstrates the impact of doppelgänger effect on ML performance and is positively correlated. The experimental results in this paper also confirm the phenomenon that in the protein sequence function prediction experiment, the sample data contains a lot of data doppelgängers under the premise that the model training results are good but the model's promotion is not good.

3.3 How to identify and minimise data doppelgängers

Although it is difficult to solve the doppelgänger effect, it is extremely necessary to solve the doppelgänger effect during the model testing and evaluation stage. In the past ten years, many solutions have been provided to solve the influence of doppelgänger effect.The identification of data doppelgängers can be solved by choosing logical methods. Given the potential of doppelgänger effects to confound, it is crucial to be able to identify the presence of data doppelgängers between training and validation sets before validation. Ordination methods (e.g., principal component analysis) or embedding methods (e.g., t-SNE), combined with scatterplots to observe the sample distribution.Then, in 2014, Q.Sheng et al. proposed the dupChecker method to identify the double body data of fingerprints[2].In 2016, L. Waldron et al proposed Another measure, the pairwise Pearson's correlation coefficient (PPCC), captures relations between sample pairs of different data sets. An anomalously high PPCC value indicates that a pair of samples constitutes PPCC data doppelgängers[3].

In the actual situation of reducing doppelgänger effects, firstly, we should use metadata as a reference to identify possible data doppelgängers, and classify them all as training set or test set, and then stratify the training set and test set during model evaluation , grouping similar data into the same level. Secondly, calculating Pearson's Correlation Coefficient (PCC) between every sample in one dataset against every sample in the other dataset. Thirdly, duplicate-oriented out lier detection. The background distribution of pairwise PCC values varies depending on the tissue assayed and the technologies used, and must be estimated for every dataset pair. Doppelgängers can be identified as outliers at the high end of the distribution of batch-corrected correlations[3]. Finally, divergent validation should be done on different independent data sets. Through the above steps, minimize the impact of data doppelgängers.

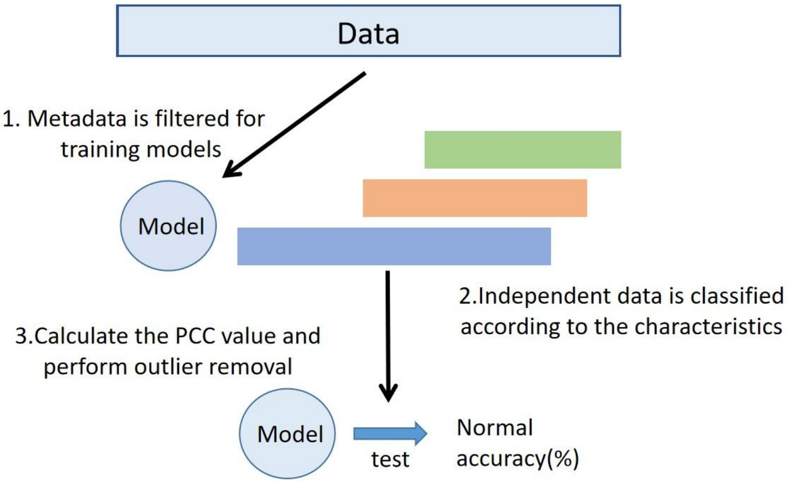
Data
1. Metadata is filtered for training models
Model
2.Independent data is classified according to the characteristics
3.Calculate the PCC value and perform outlier removal
Model
test
Normal accuracy(%)

figure 1 minimize doppelgänger effects diagram

4 Relation between doppelgänger effect and biomedical data

Meaningful and clean data, in the sense that the captured features are informative and that the data are devoid of noise or confounding effects, may yield accurate models. However, in practice, any collected data are effectively a sample, and may not be a true representation of the population under study. Sample bias, and where possible check for errors stemming from sample size, heterogeneity, noise, and confounding factors often exist in metadata, which are very common and difficult to avoid[4].

The doppelgänger effect is essentially caused by the fact that the independent sample data are too close. The emergence of this problem will have an impact on the verification part of machine learning in any field. In the process of verifying the model, neither Internal Validation nor External Validation can directly solve the data doppelgängers directly caused by the original data, resulting in model training results only Show the problem of good prediction accuracy, so data doppelgängers can exist in any field. And because biomedical raw data will inevitably generate a large number of similar independent data, the doppelgänger effect has a great impact on the biomedical field and cannot be ignored.

5 Interesting example in other data types

Whole genomes of cancer specimens are frequently analyzed, however further manual inspection of expression data, clinical annotations, etc. often results in highly replicated independent samples. In 2016, L. Waldron et al. studied ovarian cancer, breast cancer, bladder cancer and colorectal cancer and cell line databases, and found that doppelgänger effect was found in more than general experiments. For example, among 1467 breast cancer gene expression profiles, 59 samples showed doppelgänger effect. In the Ovarian Cancer Database, 17% of records were identified as non-unique, including from different datasets from the same institution. This finding has important implications for biology, as previous work on identifying duplicate microarray patterns was limited to matching identical raw data files, but this type of duplicate data is often overlooked but is also a contribution to the field of cancer. An important breakthrough in progress.

Reference

[1] Wang, L.R., Wong, L. and Goh, W.W.B., 2021. How doppelgänger effects in biomedical data confound machine learning. Drug Discovery Today.

[2] Q. Sheng, Y. Shyr, X. Chen, DupChecker: a bioconductor package for checking highthroughput genomic data redundancy in metaanalysis, BMC Bioinform 15 (2014) 323.

[3] L. Waldron, M. Riester, M. Ramos, G. Parmigiani, M.Birrer, The Doppelgänger effect: hidden duplicates in databases of transcriptome profiles, J Natl Cancer Inst108 (2016) djw146.

[4] S.Y. Ho, K. Phua, L. Wong, W.W.B. Goh, Extensions of the external validation for checking learned model interpretability and generalizability, Patterns 1 (2020) 100129.