u201812303@alumni.hust.edu.cn
https://github.com/Zealous0/Test-Question.git

Report

Original article: Wang, L. R., Wong, L., & Goh, W. W. B. (2022). How doppelgänger effects in biomedical data confound machine learning. Drug discovery today, 27(3), 678–685. https://doi.org/10.1016/j.drudis.2021.10.017

1. Introduction

Today, machine learning (ML) models are increasingly used in various fields and are prevalent in biomedical engineering, where drug development fields can use ML models, can list better drug candidates (targets) faster and reduce the time spent discovering targets and testing drugs. It can also identify existing FDA-approved drugs for the treatment of other diseases, i.e., for drug repurposing, greatly reducing the cost of drug development. In the field of medical image imaging such as PET, CT, and MRI, PET/CT image histology feature pictures can be extracted for predictive diagnosis of cancer characteristics using ML methods, such as predicting the mutation status of multiple oncogenes [1]. In PET image reconstruction, machine learning techniques have been applied to estimate the location and arrival time of high-energy photons. In quantitative image reconstruction, machine learning has been used to estimate various correction factors, including scatter events and attenuated images, as well as to reduce statistical noise in reconstructed images. Machine learning can provide alternatives to reconstructions that are less time-consuming [2]. Applying ML methods to PET imaging can also improve the resolution of the images, thus increasing their quantitative accuracy [3].

In machine learning approaches, the data are usually divided into training and test datasets and cross-validated to evaluate the model. Whereas, if independently derived datasets are very similar to each other, data doppelgängers are generated and the same dataset causes the model to perform well no matter how it is trained, and further, data doppelgängers lead to functional doppelgänger (confounding ML results), known as the doppelgänger effect. This effect was shown to be present also in the case of large data sets and similar data sets are not characterized in the whole and cannot be distinguished. While biomedical data, due to its large data volume and inflationary effects, the doppelgänger effect is prevalent in biomedical engineering, which can lead to inflated ML performance results, and poorly trained models may still perform well and not match the real situation. Thus hoping to mitigate the doppelgänger effect, the paper proposes some suggested methods to identify data doppelgängers, such as suggesting identifying doppelgängers data before training-validation dataset separation. However, most of the methods are not generalized or robust enough, and the search for methods to better mitigate the doppelgängers effect is considered.

2.Methods

A typical example is an existing chromatin interaction prediction system that is evaluated on test sets that have a high degree of similarity to the training set, yielding a data doppelgänger effect. Proteins with similar sequences in this system are inferred to be from the same ancestral protein and thus inherit the function of that ancestral protein (i.e., the two proteins are presumed to be functionally similar). This application is correct in most cases (the case of data doppelgängers), giving a false impression of a highly accurate prediction. However, upon further examination, the article analyzes that this approach will not correctly predict the function of proteins with less similar sequences but similar functions, producing a functional doppelgänger effect. Considering the potential confounding effect of the doppelgänger effect, it is crucial to be able to identify the presence of data doppelgängers between the training and validation sets prior to validation.

For this case, one logical approach proposed for biological data doppelgänger identification is to use ranking methods (e.g., principal component analysis) or embedding methods (e.g., t-SNE) coupled with scatter plots to see how the samples are distributed in the reduced dimensional space. However, the article finds that this approach is not feasible because the data doppelgängers are distinguishable in the reduced dimensional space.

An earlier method also proposed is dup checker, which identifies duplicate samples by comparing the MD5 fingerprints of their CEL files The same MD5 fingerprint indicates that the samples are duplicates (essentially duplicates, thus indicating a leakage problem). Thus, doppChecker does not detect real data doppelgängers, which are independently derived samples that are accidentally similar.

Another early proposed approach is the paired Pearson correlation coefficient (PPCC), which captures the relationship between sample pairs from different datasets An unusually high PPCC value indicates that a pair of samples constitutes a PPCC data doppelgänger But the main limitation of the original PPPC paper is that it never ends up confusing PPCC data doppelgängers with their ML task's ability (i.e., to have functional effects and therefore act as functional doppelgängers). However, PPCC is reasonable as a quantitative measure of data doppelgängers. In the article, PPCC was constructed in different scenarios by using renal cell carcinoma (RCC) proteomics data obtained from the NetProt software library by Guo et al. The results showed that PPCC values were generally high for the same tissue pairs and low for different tissue pairs, and the results indicated that PPCC has significant discrimination value and can be used as an evaluation of data doppelgängers effect as a quantitative indicator for evaluating data doppelgängers effect.

Based on the proposed PPCC evaluation index, an attempt was made to demonstrate that the doppelgängers effect has a confounding effect on ML results. My article explores the effect of doppelgängers on validation accuracy in different randomly trained classifiers. The results demonstrate that the presence of PPCC data doppelgängers in the training and validation data inflates ML performance even if these features are randomly selected, i.e., the model performs poorly during validation. Moreover, the more doppelgänger pairs represented in the training and validation sets, the more ML performance is inflated. When the validation accuracy is such that the data for all correctly trained models are stratified into PPCC data doppelgängers and non-PPCC data doppelgängers layers, all ML models exhibit higher performance on PPCC data doppelgängers than on non-PPCC data. This suggests a dose-dependent relationship between the number of PPCC data doppelgängers and the magnitude of the doppelgänger effect.

On this validation result, a possible method to avoid the doppelgänger effect is proposed by constraining the PPCC data doppelgängers to either the training set or the validation set. However, in the former case, the size of the training set is fixed, which can lead to a lack of knowledge for the model to generalize well. In the latter case, the predictions of the doppelgängers are either correct or incorrect.

3、Concliusion and Suggestions

Although some proposed methods for identifying data doppelgängers exist, most of them are not generalized or robust enough. The article suggests 3 ideas for a solution.

The first one is to use metadata as a guide for careful cross-checking. Here, negative and positive cases are constructed by using metadata from the RCC. This allows us to predict the range of PPCC scores for scenarios where doppelgängers are not present (different classes; negative cases) as well as cases where there is leakage (based on duplicates of the same patient and the same class; positive cases). The plausible data doppelgängers of interest are samples from the same class but different patients. With this information from the metadata, we were able to identify potential doppelgängers and effectively classify them all into training sets or validation sets preventing doppelgängers effects and allowing a relatively more objective evaluation of ML performance. Likewise, technical replicates generated from the same sample should be treated similarly. This recommendation is similar to guidelines in bioinformatics, which suggest that when ML models are trained with data from biological sequences, researchers should ensure that the training and test samples are not duplicates or highly similar samples.

The second recommendation is to perform data stratification. Instead of evaluating model performance on the entire test data, the data can be stratified into layers with different degrees of similarity (e.g., PPCC data doppelgängers and non-PPCC data doppelgängers, and model performance on each layer can be evaluated separately). More importantly, strata with poor model performance can pinpoint gaps in the classifier. In RCC, the non-ppcc doppelgängers used for stratification performance assessment also happened to be papillary RCC samples. Considering that the proportion of kidney cancer cells in each tissue is known (papillary RCC includes 10% of kidney cancer cells),2 the poor performance of the classifier on papillary RCC suggests that this 10% sample of kidney cancer cells is the weak point of our classifier will need further improvement.

The third recommendation is to perform very robust independent validation checks involving as many datasets as possible (divergence validation). Although not a direct hedge against data doppelgängers, the different validation techniques can inform the objectivity of the classifier. It also informs the generalizability of the model despite the possible presence of data doppelgängers in the training set. Future research can explore other functional doppelgänger identification methods that do not rely heavily on metadata. In this approach, functional doppelgängers can be identified directly). Further combining this approach with PPCC may subsequently be able to identify test set samples from the training set of the doppelgänger part.

The article shows the prevalence of doppelgänger effects in biomedical data, demonstrates how doppelgängers arise and provides evidence for their confounding effect. To mitigate the doppelgänger effect, the article suggests identifying doppelgängers data before training-validation separation, and based on this, 3 feasible suggestions are given.

References

[1]Chang, Cheng et al. “A Machine Learning Model Based on PET/CT Radiomics and Clinical Characteristics Predicts ALK Rearrangement Status in Lung Adenocarcinoma.” Frontiers in oncology vol. 11 603882. 2 Mar. 2021, doi:10.3389/fonc.2021.603882

[2] K. Gong, E. Berg, S. R. Cherry and J. Qi, "Machine Learning in PET: From Photon Detection to Quantitative Image Reconstruction," in Proceedings of the IEEE, vol. 108, no. 1, pp. 51-68, Jan. 2020, doi: 10.1109/JPROC.2019.2936809.

[3] Song TA, Chowdhury SR, Yang F, Dutta J. Super-Resolution PET Imaging Using Convolutional Neural Networks. IEEE Trans Comput Imaging. 2020;6:518-528. doi:10.1109/tci.2020.2964229