1. Introduction of Doppelgänger effects

Data Doppelgängers are defined as the samples which are independent but similar[1]. The data needs to be diversified. If they are similar data, model complexity is too low and too few data features learned by the machine. Hence, it may be easy to cause underfitting. Data doppelgängers may result in independent derived training sets and validation sets performing well regardless the quality of training, thereby confusing the performance of machine learning (ML) models[2], which is called doppelgänger effects.

It can be commonly found in biomedical data. For example, F. Gao et.al[3] found the inflation of ML when they evaluate the existing chromatin interaction prediction system[3]. What's more, data doppelgängers are also present in protein function prediction, proteins with similar sequences are inferred to be descended from the same ancestor protein.[1] Also, several independent studies have noted the presence of confounding similarities in RNA families (Szikszai et al., 2022)[4], or shared ancestry (Greener et al., 2022)])[5]. F Cao et al. focus on identifying PPCC DDs in two RNA-Seq data sets, lymph_lung and large_upper[3]. In my perspective, doppelgänger effects can also be found in other professional directions as long as independent but similar data appear.

2. Ways to check doppelgänger effects

As we can infer from "How doppelgänger effects in biomedical data confound machine learning"[1], having data doppelgänger does not necessarily result in doppelgänger effects. Only when data doppelgänger exists, and the classifier shows false excellent performance because of this, can we call this doppelgänger effect. For example, even if some of the samples in the two data sets are data doppelgänger, but the size of samples is large enough, the data doppelgänger is present but powerless, and the classifier can still learn the data well and obtain real and good performance, so that even if the data doppelgänger appears, there is no doppelgänger effect, and the performance of the classifier is unaffected. Then, data doppelgänger which leads to doppelgänger effects is called functional data doppelgänger.

Hence, we can check doppelgänger effects from two steps. First, data preprocessing stage, identify the presence of data doppelgänger before training the classifier. Second, evaluation on potential functional data doppelgängers, check whether data doppelgänger leads to doppelgänger effects.

2.1 Data preprocessing stage

The first step is detailed in "How doppelgänger effects in biomedical data confound machine learning". As ordination methods(e.g., PCA) or embedding methods (e.g., t-SNE) are unfeasible because of the reduced-dimensional space and dupChecker because of the data leakage , Li Rong Wang et.al[1] use the pairwise Pearson's correlation coefficient (PPCC) to identify the potential functional data doppelgänger by using the data from RCC, following five steps below.

1. Batch correction, construct benchmark scenario, data are separated into negative, valid and positive cases;
2. Calculate PPCC between samples of different datasets;
3. Group sample pairs by similarity of different classes;
4. Calculate Cut Off, calculate the maximum PPCC value for all negative sample pairs, denoted as cut off (ignore outliers);
5. Valid sample pairs with PPCC values bigger than cut off are data doppelgängers[1].

we may construct a file describing the samples in each training-validation pair. Training-validation pairs should be chosen strategically to have incrementally increasing numbers of PPCC data doppelgängers s between the training and validation sets. This allows us to observe the inflationary effects of the PPCC data doppelgängers s easily[3].

Also, other correlation metrics such as the Spearman Rank correlation coefficient and Kendall Rank correlation coefficient can have the same effect as PPCC and be applied into identifying data doppelgängers.

Take the PPCC of RCC data to set as an example to analyze (Fig.1)

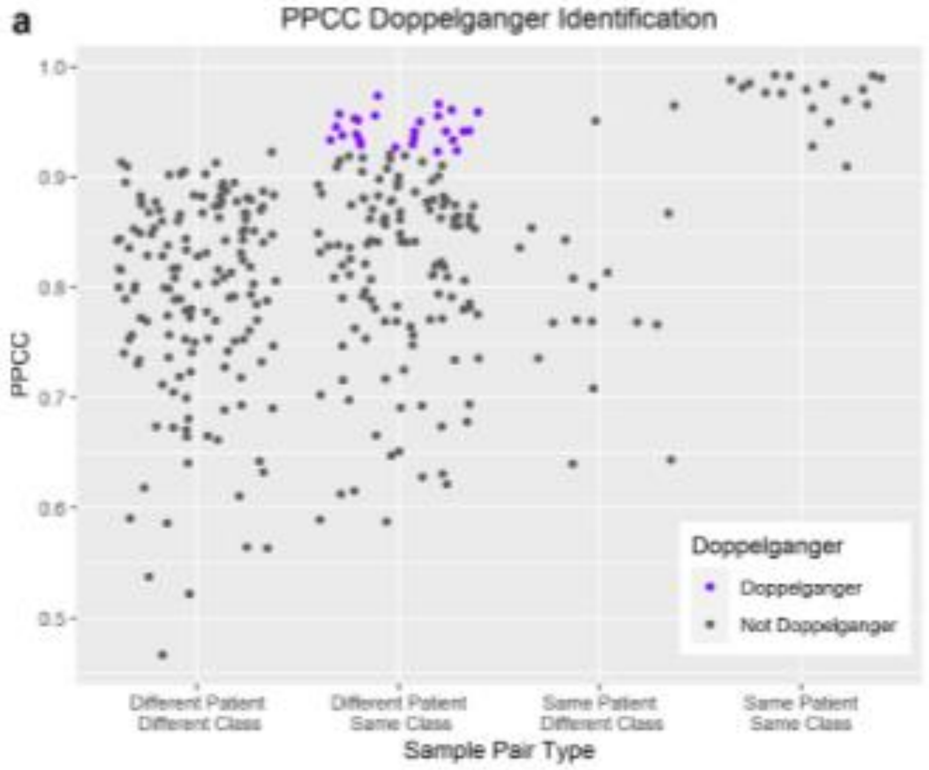

Fig.1 Distribution of pairwise Pearson’s correlation coefficients (PPCCs) across different sample pairs

As we can indicate from the distribution of pairwise Pearson's correlation coefficients (PPCCs) across different sample pairs. The data in the first and third column belong to the negative case, data in the second column belongs to the valid case and data in last column belongs to the positive case. The data in the valid case whose PPCC is bigger than all data in the negative case is defined as data doppelgänger. We can also infer that the data whose PPCC is between 0.9 and 1 in the third column may be the outliers.

2.2 Evaluation on potential functional DDs

Li Rong Wang et.al[1] use k-nearest neighbor (kNN) ,Naïve Bayes, Decision Tree and Logistic Regression models in which the training validation set with doppelgängers in validation to evaluate an identical accuracy distribution to the training-validation set with perfect leakage. By putting all the splits into the training set, the accuracy drops to 0.5, which is the expected accuracy based on the random feature training model.

About how to evaluate the potential functional data doppelgängers, in my perspective, we can start with other evaluation criteria of the classifier besides accuracy, such as recall rate, precision rate, confusion matrix, AUC value, etc. These criteria can give evaluation in more aspects, which benefits a lot on doppelgänger effect identifying.

ACC has many drawbacks as an evaluation index. When applied to imbalanced datasets, ACC provides an overly optimistic evaluation of most classifier performance[6]. It is not appropriate to only use ACC as a classifier evaluation index in some cases.

The most popular and widely used measure is the Kappa coefficient. The Kappa coefficient was developed by Cohen in 1960 to test reliability among raters[7]. In the past few decades, Kappa coefficient has also entered the field of machine learning to compare classifier performance.

The Matthews correlation coefficient MCC was first proposed by Matthews in 1975 to compare chemical structures[8]. In 1920, it was re-proposed by Baldi et al. Since then, this index has been introduced into the field of machine learning. Because the MCC takes into account all error situations, the MCC is considered more reliable than the ACC[6]. But this indicator applies only to binary classification problems.

In addition, there are AUC evaluation indicators based on ranking indicators calculated according to the ROC curve. The ROC curve is drawn on the basis of selecting two categories, with the classification error rate of one category as the abscissa and the classification accuracy of the other category as the ordinate. However, when the curves intersect, the direct observation curve cannot intuitively measure the performance of the classifier, so the AUC index is proposed. It can be seen from the origin of the index that the AUC index is also only applicable to the binary classification problem, and the AUC index takes into account the classification effect of both categories.

3. Ways to avoid doppelgänger effects

As we can learn from Li Rong Wang et.al "How doppelgänger effects in biomedical data confound machine learning"[1], there are three feasible ways to avoid doppelgänger effects.

First, use metadata as a guide for careful cross-checking. This allows us to predict PPCC score ranges and data leakage in scenarios where doppelgängers do not exist. With this information from metadata, we can identify potential doppelgängers and classify them all into training or validation sets, effectively preventing the doppelgänger effect and allowing for a relatively more objective evaluation of ML performance[1].

Second, instead of evaluating the performance of the model over the entire test data, we can layer the data into layers of different similarities to evaluate the performance of the model on each layer separately[1].

Third, perform independent verification checks involving as many data sets as possible. Although not a direct hedge against data doppelgängers, different validation techniques can inform the objectivity of the classifier. It also informs the generalization of the model, although data doppelgängers may exist in the training set[1].

References

[1] Wang, L. , L. Wong , and W. Goh . "How doppelganger effects in biomedical data confound machine learning. " Drug discovery today (2021).
[2] Ho, S. Y. , et al. "Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability." Patterns 1.8(2020):100129.
[3] F. Cao, M.J. Fullwood, Inflated performance measures in enhancer–promoter interaction-prediction methods, Nat Genet 51 (2019) 1196–1198.
[4] Szikszai M., Wise M.J., Datta A., Ward M., Mathews D. Deep learning models for RNA secondary structure prediction (probably) do not generalise across families. bioRxiv. 2022 doi: 10.1101/2022.03.21.485135. Preprint at.
[5] Greener J.G., Kandathil S.M., Moffat L., Jones D.T. A guide to machine learning for biologists. Nat. Rev. Mol. Cell Biol. 2022;23:40–55.
[6] Chicco, Davide , and G. Jurman . "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." BMC Genomics 21.1(2020).
[7]Cohen, and J. "A Coefficient of Agreement for Nominal Scales." Educational & Psychological Measurement 20.1(1960):37-46.
[8] Matthews, B. W. . "Comparison of the predicted and observed secondary structure

of T4 phage lysozyme. " Biochim Biophys Acta 405.2(1975):442-451.