How doppelgänger effects in biomedical data confound machine learning

1. Why ML

(1) an shortlist better drug candidates faster, reducing time spent on discovery and testing.

(2) classification models based on ML can predict new drug-disease interactions and possible adverse drug reactions.

2. Data doppelgängers,doppelgänger effects and functional doppelgängers

data doppelgängers: they are data where training and validation sets are highly similar and models trained and validated on data doppelgängers might perform well regardless of the quality of training. The model that appears to be well trained on similar data will give out similar outcomes when meeting similar inputs. However, when meeting the dissimilar input which should have a similar output, the model will adversely give out a completely different outcome. That is because of the similarity between training data and validation data. The model seems to be well trained but in fact not since it has only been fed with the same kind of inputs whose labels are also similar. Then, the variables in the model only remembered that same kind of features(vectors). Finally, I guess, that model only remembers those feature vectors in that kind of data. However, when the feature vectors are totally different, the outcome will also be different. In this manner, the cross-validation technique will be useless.

doppelgänger effect happens when a classifier falsely performs well because of the presence of data doppelgängers.

Functional doppelgängers are data doppelgängers that also generate a doppelgängers effect (confounding ML outcomes).(What really needs to be identified out)

Extra point:

I will give an example of the doppelgängers effect in imaging classification below.

For instance, when we train a classifier to recognize a cat, we only train on images from one kind of cat such as American Shorthair cat combined with other animals images like dogs and pigs. Then, the validation set is also generated from American shorthair cat images and other kind of animals' images. In this situation, the model would perform well because it precisely remembers the feature vectors of American Shorthair cat. However, when met a African cat, the model could recognize it as a dog or anything else since the feature vectors of African cat are different from American Shorthair cat. This is caused by the doppelgängers effect. By analyzing the model, I think the variables in the model were not well trained because of lacking data. Maybe, if the model is too small, it could not have the ability to fully learn many features. Assuming the model is big enough to learn thoroughly, it might also fail to do so since it has never tried to distinguish the small difference between different kinds of cats due to being short of similar cat images.

If there are various kinds of cat images in the training set, I believe the variables in the model will learn the small difference between different cats and distinguish them from other kinds of animals.

To sum up, I suppose the doppelgängers effects not only appear in biomedical data, but also happen in many other kinds of data such as images and videos. I think this effect will occur when lacking the samples that differ from the targets but have the same outcomes no matter what kind of data.
Therefore, the solution to avoid it is to supplement more relevant data that differ from the target input but has the same classification.

However, I know it is very hard for staff to obtain more biomedical data from more patients. Thus, I think maybe we can use the few-shot learning method which only needs to use a few data to train a really useful model. As for few-shot learning, we can generate some 'fake' data by learning the feature from another set of true data. For instance, when there are ample kinds of cat images but only one kind of dog images, we can generate some 'fake' dog images by learning from those cat images. These 'fake' images can seem to be real and compensate for the lack of similar data.

3. Goal

Prove the confounding effects of doppelgängers effect and understand better the level of similarity between suspected functional doppelgängers and the acceptance proportion of functional doppelgängers in the validation set. If possible, try to solve this question.

4. Related work

(1) Found of data doppelgängers

a. Cao and Fullwood found the data doppelgängers in chromatin interaction prediction systems;
b. The author found that the abductive reasoning would not precisely predict functions for proteins with less similar sequences but similar functions due to data doppelgängers. The quantitative structure-activity relationship(QSAR) models also have the same problem while predicting the biological activities of molecules from their structural properties.

(2) Identification of data doppelgängers

a. Principal component analysis or embedding method, coupled with scatterplots, to see how samples are distributed in reduced-dimensional space. Weakness: does not always work because data doppelgängers are not necessarily distinguishable in reduced-dimensional space.
b. DupChecker identifies duplicate samples by comparing the MD5 finger-prints of their CEL files. Weakness: can not detect true data doppelgängers that are independently derived samples that are similar by chance.
c. Pairwise Pearson's correlation coefficient (PPCC) captures relations between sample pairs of different data sets, while an anomalously high PPCC value indicates that a pair of samples constitutes PPCC data doppelgängers. Weakness: does not make a link

between PPCC data doppelgängers and their ability to confound ML tasks. And their reported doppelgängers were in fact the result of leakage (between sample replicates), so it could not constitute true data doppelgängers.

5. How to differentiate poorly trained models from well-trained ones

Take molecules activities prediction as an example. The author tested their performance on similar molecules with different activities(SAR paradox). In this experiment, a well-trained model would still perform well while the poorly trained one would fail to tell the difference.

6. New method to identify data doppelgängers

(1) data set: a renal cell carcinoma(RCC) benchmark data set.

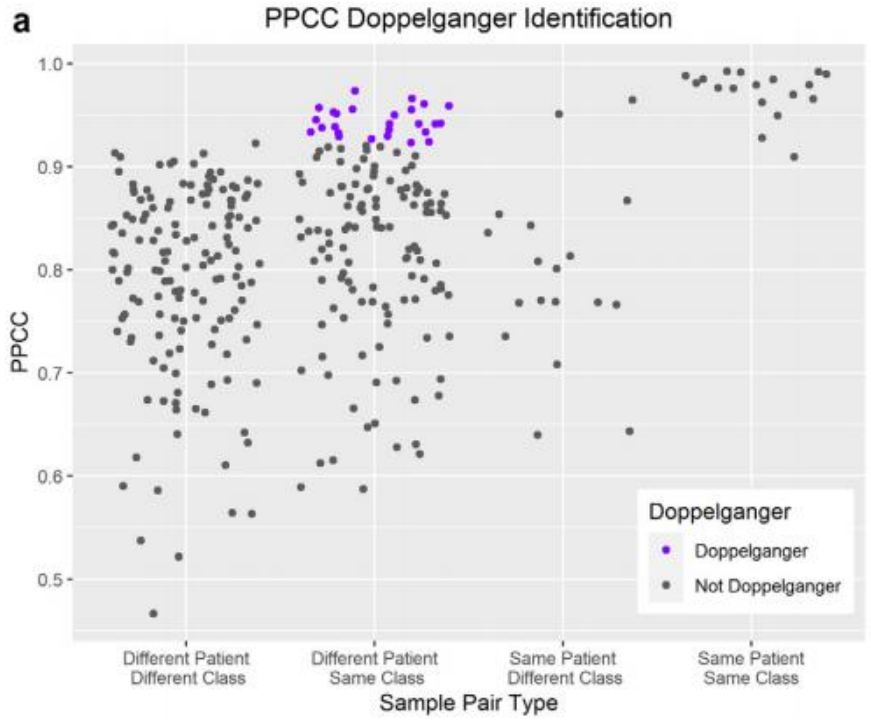
Three types of cases constructed from two batches of the RCC data set:

Negative cases: consist of sample pairs from different class labels(no doppelgängers)

Positive cases: consist of sample pairs from same sample(not doppelgängers)

Valid cases: consist of sample pairs assigned to the same class label but from different samples.

(2) The identification of PPCC data doppelgängers is based on the PPCC distribution of the valid scenario against the negative and positive scenarios.



(The purple points in the figure represent the doppelgängers because they come from different samples while have high PPCC values)

(3) Trails result:

a. PPCC distributions on the valid scenario exist as a wide continuum, without obvious breaks, suggesting that data doppelgängers exist naturally as part of the similarity spectrum between samples.

b. PPCC values for tissue pairs come from the same samples and different samples are high.
c. PPCC distributions are lower when comparing different tissue pairs.
d. The performance of validation accuracy of PPCC data doppelgängers was higher than that of non-PPCC data doppelgängers.
e. This result proved that the presence of PPCC data doppelgängers in both training and validation data does inflate ML performance, regardless of what training and validation data they use and what model they train. Meanwhile, the more doppelgängers pairs represented in both training and validation sets, the more inflated the ML performance. Moreover, it also confirmed that PPCC data doppelgängers could act as functional doppelgängers and produce inflationary effects similar to data leakage(sample replicates) when using kNN model or naive bayes model to train.

7. Method to avoid doppelgängers

(1) Placing all doppelgängers in the training set. However, when the size of training set is fixed, the other kind of data that are less similar would not be included in the training set, then the model would not be well trained because of lacking these data.
(2) Constraining the PPCC data doppelgängers to either the training or validation set. While in this situation, if the doppelgängers were divided into training set, they would be well trained and be predicted correctly. However, if they were divided into validation set, the model may lack the knowledge to predict them.
(3) Removing all PPCC data doppelgängers. This might reduce the data to an unusual size which may hinder the training of the model.
All of the methods below need prior knowledge to locate those data doppelgängers which is very difficult.
(4) Removing variables contributing strongly toward data doppelgängers effects. This method is useful in theory but so hard to operate because the effect can not be simply explained by a subset of highly correlated variables.

8. Recommendation

(1) Performing careful cross-checks and identifying potential doppelgängers using meta-data as a guide. Then, we can execute method(2). When constructing the data, we should control the similarity of training and test samples.
(2) Performing data stratification. Through testing on divided test data, we can not only avoid the data doppelgängers effect, but also know the weakness of our classifier from those classes with poor performance.
(3) Involving as many data sets as possible when testing. If trained on enough data, the model will be qualified to identify data doppelgängers.

9. Other possible orientation:

(1) Find a more straightforward functional doppelgängers identification method without relying on meta-data;
(2) Look for subsets of a validation set that are predicted correctly regardless of the ML method used. Then, keep their doppelgängers partner out of the training set.