Literature Report

Title: How doppelgänger effects in biomedical data confound machine learning
Author: Li Rong Wang , Limsoon Wong , Wilson Wen Bin Goh.
Journal: *Drug Discovery Today*
Published year: 2022

Summary of Core Work:

The authors proposed a method for identifying data doppelgänger in data sets based on meta-data and pairwise Pearson's correlation coefficient (PPCC) and demonstrated that the presence of PPCC data doppelgängers in both training and validation data would inflate ML performance.

Introduction

Machine Learning (ML) can increase the efficiency of drug discovery. Classifiers have been used for the prediction of new drug–disease interactions and possible adverse drug reactions. In the field of ML, two independent data sets, training set and test set, should be used to assess the performance of a classifier. However, due to the data doppelgängers, the assumption that the two data sets are independent from each other is not always true. When the training set and test set data are very similar for some reason, the similar data is called data doppelgängers. The data doppelgängers may improve the performance of the ML model regardless of the quality of training. This effect is called doppelgänger effects. It is still uncommon to check the data doppelgängers before training and testing ML models. In this paper, authors want to measure the level of similarity and the acceptable proportion of functional doppelgängers and propose improved methods for doppelgänger identification.

Abundance of data doppelgängers in biological data

Many works have revealed the abundance of doppelgänger data in biological data. In protein function prediction, proteins with similar sequences are presumed to be similar in function. This abductive reasoning is true in most cases (cases of data doppelgängers). However, this approach would be unable to correctly predict functions for proteins with less similar sequences but similar functions. QSAR models, widely used in drug discovery, assume that structurally similar molecules have similar activities. But these models can not differentiate the small structure variations that can cause significant changes in activity. Doppelgänger data will confound model validation by making poorly trained models, which have uninformative structural properties, still work well on the test data.

Identification of data doppelgängers

Authors found that doppelgängers are not necessarily distinguishable in reduced-dimensional space. The dupChecker method can only find the data leakage but cannot detect true data doppelgängers that are from independent datasets but similar by chance. Pairwise Pearson's correlation coefficient (PPCC) has been used to identify the data doppelgängers. However, the original paper cannot conclusively make a link between PPCC data doppelgängers and their ability to confound ML tasks. Therefore, the author adopted a method based on meta-data and PPCC.

They first constructed a benchmark scenario. Two sets of proteomics data taken from the NetProt software library. Each data set contains proteomics data from many patients. Each patient has two types of data: data from tumor tissue and data from normal tissue. In order to construct the positive group, some patients' data are duplicated and added to another data set. These patients' data appeared in both two datasets. Two samples from different datasets constitute a sample pair. There are 4 types of sample pairs: (1) Same patient same class. This type is the positive group, since they are identical and have the strongest correlation. This type is obvious data leakage and not considered as data doppelgängers. (2) Different patient same class. They are the valid case where we need to find the doppelgängers. (3) different patient different class (4) same patient different class. These two types are negative groups since the data belong to two different classes.

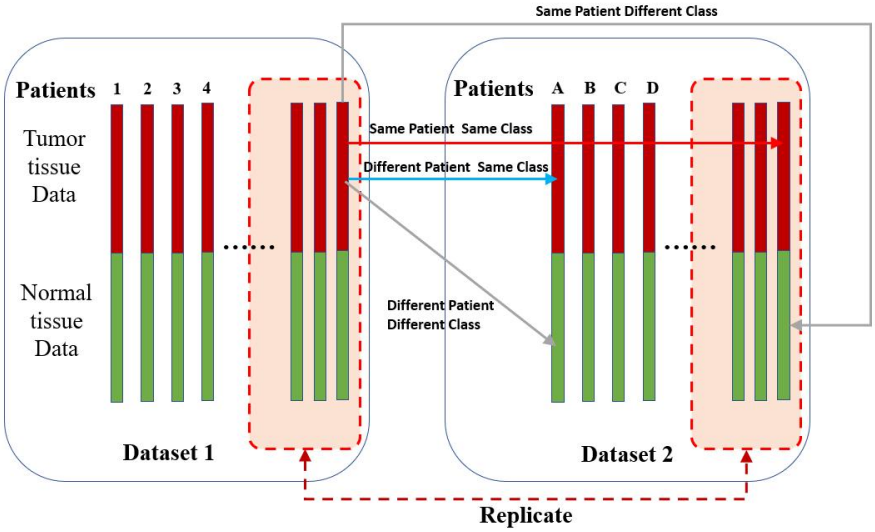

Figure 1: Benchmark scenarios for identification of data doppelgängers

The purpose of constructing a benchmark scenario is to obtain a meta-data to judge the data doppelgänger. In other words, the positive group and negative group are used to obtain a threshold of PPCC value. The data pair whose PPCC value is higher than this threshold can be regarded as data doppelgängers. The identification result is shown in Figure 2. Authors found that half of samples in the dataset are PPCC data doppelgängers with at least one other sample, which reveals the prevalence of data doppelgängers in biological data.

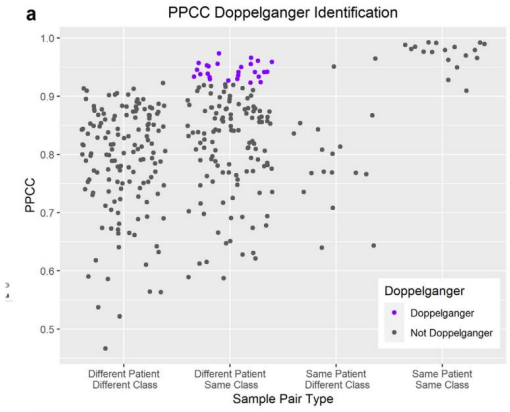

Figure 2: The identification result

Confounding effects of PPCC data doppelgängers

In order to determine whether the PPCC data doppelgängers can affect the performance of ML models on the validation set, authors use the dataset to train different ML models and test their performance on designed validation sets. 10 randomly generated feature sets are used to train the ML model. 'Top 10% Variance' feature set and 'Lowest 10% Variance' feature set are used to represent the good training and poor training model respectively. Five different validation sets are designed as having 0,2,4,6,8 doppelgängers respectively. Besides, another dataset containing 8 identical data from the training set is designed as a perfect leakage group. Binomial group is used as a negative control since it is equal in performance to a random coin toss.

Four types of ML models are tested. K-Nearest Neighbours, Naïve Bayes, Logistic Regression and Decision Tree models. The results show that the PPCC data doppelgängers can inflate the performance of some ML models. The inflationary effects are obvious on the KNN model and Naïve Bayes model, but not clear on the Logistic Regression and Decision Tree models. In KNN case, the accuracy increases with the number of doppelgängers. In fact, when the validation set is filled with the doppelgängers (8-doppel), the performance of all ML models are almost perfect even if many of them are trained randomly. However, when the validation set has no doppelgängers, the performance of randomly trained ML models is very poor. The conclusion is that the PPCC data doppelgängers and data leakage have similar confounding effects on ML models.

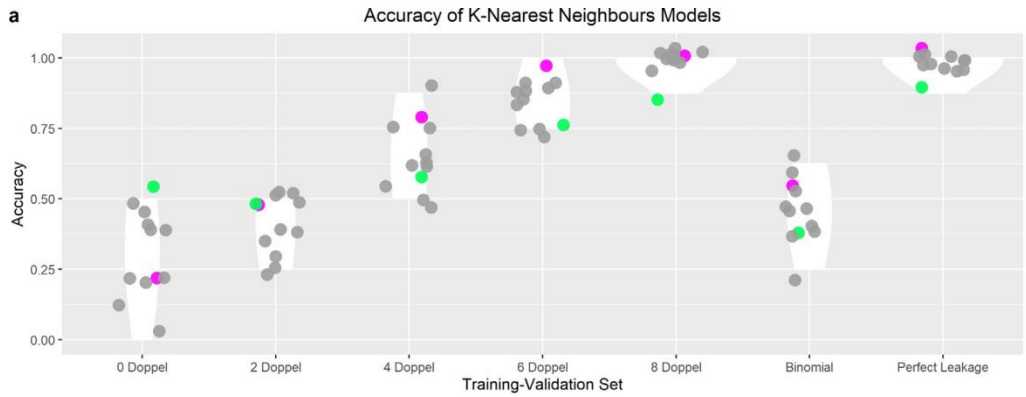

Figure 3: The confounding effects

Recommendations

The authors' first recommendation is to perform careful cross-checks before training models. In this paper, they use meta-data as a guide to distinguish the data doppelgängers. The plausible data doppelgängers are samples arising from the same class but different patients. However, authors also anticipate other identification methods that do not rely heavily on meta-data.

The second recommendation is to perform data stratification. If people can stratify data into strata of different similarities and each strata coincides with a known proportion of real-world population, then the performance of the ML model in the real world can be predicted by testing the model on different data strata.

The third recommendation is to perform extremely robust independent validation checks. The dataset should be as large as possible, which can reduce the inflationary effect brought by data doppelgängers.

[1] Whether you think doppelganger effects are unique to biomedical data?

Doppelganger effects are very common in data science. The face recognition technology is an example. If a validation set contains only humans and objects, then an algorithm based on recognizing only eyes will have perfect performance on the validation set. But it will recognize a dog as a person in the real world.

[2] How you think it can be avoided in the practice and development of machine learning models for health and medical science.

[3] propose interesting and useful ways of avoiding or checking for doppelganger effects.

The doppelganger effects arise from the lack of representativeness of the validation set. Therefore, a direct way of avoiding it is to expand the test set. A large, representative test set can solve this problem. It is not a good idea to constrain all the data doppelgängers to either the training or validation set, because in the real task, the algorithm will definitely run into a sample similar to the training data. In my opinion, the best way is to construct a representative test set reflecting the real situation truthfully. This would be a consortium level effort. For individual researchers, we should use different datasets to validate our algorithm. In addition to improving the reliability of the validation, interdisciplinary cooperation can assist model training by introducing more prior knowledge to remove irrelevant information that affects model training.

[4] You can find interesting examples in other data types e.g. imaging, gene sequencing, metabonomics.

The doppelganger effects are very common in AI imaging diagnosis. AI diagnosis usually shows a high accuracy in validation set. However, many of them do not have good clinical performance in the real world. Many developers of these AI diagnosis algorithms use images from the same dataset. When they are applied in a different hospital, or use a different imaging machine, or change the dataset, the accuracy will be totally different. I think the reason behind this is the doppelgänger effect.

[5] Demonstrate clear understanding on how these doppelganger effects emerge from a quantitative angle

Firstly, I want to share my opinion on why the inflationary effects are more obvious on KNN model and Naïve Bayes model. **The reason is that KNN model and Naïve Bayes model are not really 'trained'. They do not generate a 'model'**. In KNN model, we just calculate the distance between the training data and test data. In Naïve Bayes model, we calculate the posteriori probability of test data based on the training data. Both of them do not process the training data and extract the information from training set. It is like that they are just comparing the two datasets. Therefore, they are more susceptible to the doppelganger effects. In contrast, Logistic Regression and Decision Tree models are "trained" and try to extract the information from training set.

The random feature sets cannot give them enough information to construct a robust model. Therefore, the existence of doppelgangers cannot improve the accuracy significantly.

In my opinion, the fundamental reason for the doppelganger effects is that **some poor models that make judgements based on unrelated information can still perform well on the validation set because of the insufficient representativeness of the validation set.** This is like a student who sees the final exam paper in advance and memorizes the question order and the corresponding answer. He is able to get high marks without understanding questions.