Doppelganger effects in Machine Learning

Written by Liu Sinuo
db92311@um.edu.mo

I. Data Doppelganger and Doppelganger effects

Machine Learning (ML) models have been increasingly used in biomedical area to speed up drug development. These models have more advantages than traditional method, more accurate and ;faster. Several ML-identifified drugst have also advanced into clinical trials. Given the drug testing, it is important to train and test these classifiers (trained models) properly to determine suitable drug candidates. However, in the process of training and testing, doppelganger effects may occur.

Data doppelgangers are when samples appear similar across their measurements (Wang et al, 2021). The samples may have similar molecules, similar actives and similar features. When independently derived data are very similar to each other, the data doppelganger appears, which makes the model perform perfectly no matter how well they are trained. Due to data doppelganger, cross - validation results are unreliable regardless of the quality of training. When a model (classifier) performs well wrongly because of the data doppelganger, there is an observed doppelganger effect. However, not all data doppelgangers will cause a doppelganger effect. Data doppelgangers that generate a dop pelganger effect are termed functional doppelgangers (Wang et al, 2021).

Data doppelgangers and doppelganger effects are quite common in biomedical data. Data doppelgangers have been observed in established fields of bioinformatics. For instance, in the prediction of protein function, proteins with similar sequences are inferred to be the descendants of the same ancestor protein, thus inheriting the function of the ancestor (Wang et al, 2021). However, after more in-depth examination, Wang realized that this method would not be able to correctly predict the function of proteins with dissimilar sequences but similar functions, such as twilightzone homologs and enzymes that are dissimilar in sequence overall but with similar active site residues. The naive application results are correct in most cases due to data doppelgangers.

There is another example in drug discovery. The quantitative structure-activity relationship (QSAR) model is used to predict the biological activity of molecules from their structural properties (Wang et al, 2021). The QSAR model assumes that molecules with similar structures have similar activities. Because of data doppelgangers, this assumption is correct in most cases. Some poorly trained models with uninformative structural attributes may still perform well.

From a quantitive prospect, doppelganger effects refer to the training dataset is highly similar to test dataset, regardless of whether it is random selected based on time or space, the validation results are always very accurate, that is, the ideal but not exist phenomenon.

II. Universality of doppelganger effects

The data doppelganger and doppelganger effects are not unique to biomedical data. The effect is real and persistent even unseen, but it is most evident through data leaks and exposures, as well as through unintended consequences for individuals in physical and virtual spaces. Doppelganger effect not only is negative, but also it can cause positive results.

The word “Doppelganger” was drawn from a German folklore. Doppelganger originally refers to two people who are the same in appearance (not biologically related). The doppelganger method was first developed by Nate Silver, a statistician and political forecaster, who used it to predict the future performance of baseball players. Silver realized that instead of trying to map a player's performance to a general career trajectory curve, it would be better to find the past player who was statistically most similar to that player.

The doppelganger effect is a series of effects, including the ability of the database to copy identities from a collection of facts and details about natural persons. The result is refined and clarified as an orderly singularity that seems legal, regardless of how wrong it may be: the misidentification of an airline passenger places them on the no-fly list or a financially-established university professor is refused a big-box store credit card (Robinson, 2008). The doppelganger effect can be found in a variety of data, including building, climate, sports, astronomical phenomena, and personal information.

Images can also have doppelganger effects. In facial recognition systems, the doppelganger usually increases the likelihood of mismatching. The example is shown in figure 1.

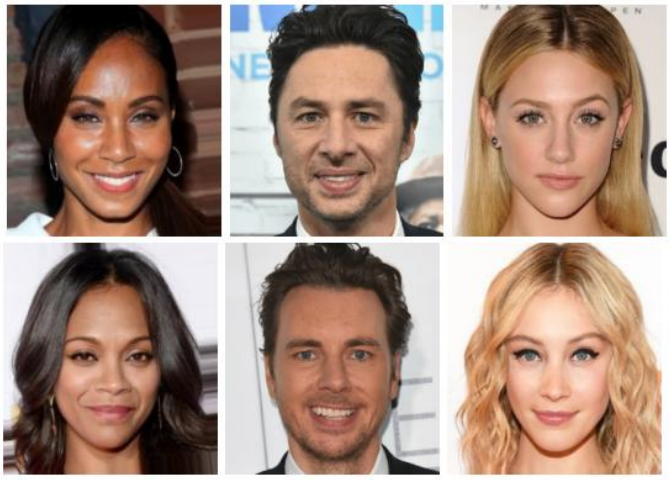


Figure 1. Image doppelgangers

In addition to demographic features, doppelgangers often share facial features, such as facial shapes. Moreover, some facial attributes can be further changed to achieve greater similarity to the target object, such as by using hairstyle or makeup. This effect has been confirmed by several researchers, which can lead to serious risks in a variety of situations, such as blacklist checks, in which innocent people may have a better chance of matching similar-looking people on the list.

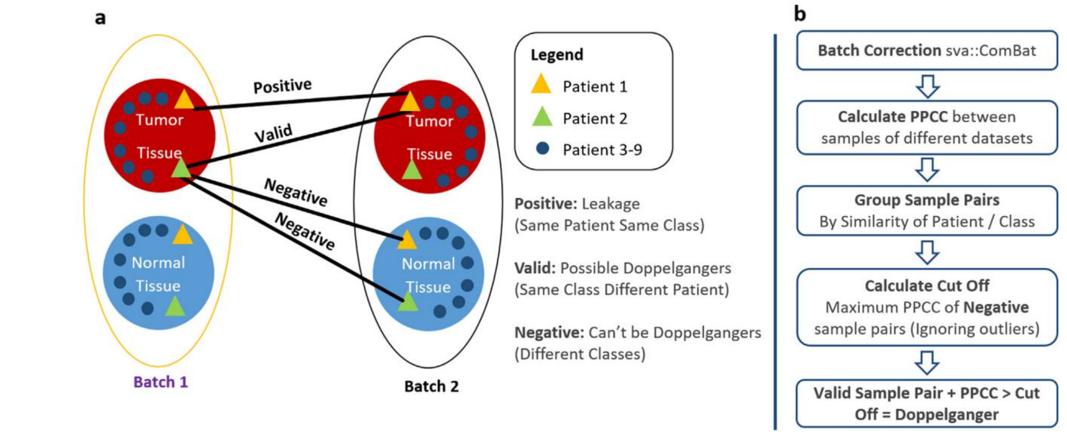
The doppelganger effect has not only disadvantages, but also advantages. For example, doppelganger effect can be used to find features that can replace some extinct creatures and record their characteristics for research.

III. Identify and Avoid Doppelganger Effects

Considering the potential problems of doppelganger effects, it is essential to identify and avoid the existence of doppelganger effects between training sets and test sets before validation, especially in the practice and development of ML models for health and medical science. There are several methods below:

- **Pairwise Pearson's correlation coefficient (PPCC) doppelganger identification**

Wang's team has used Pairwise Pearson's correlation coefficient (PPCC) doppelganger identification method to find data doppelgangers. They used renal cell carcinoma (RCC) proteomics data as benchmark scenarios. Doppelgangers are not allowed in negative cases by constructing sample pairs assigned to the same class label but from different samples, while allowed in valid cases. These effects can then be compared with positive cases which constructed by replication generated by the same sample (figure 1A). Then PPCC data doppelgangers can be identified according to the PPCC distribution of effective scenarios against negative and positive scenarios.


a
Tumor Tissue
Normal Tissue
Batch 1
Positive
Valid
Negative
Negative
Tumor Tissue
Normal Tissue
Batch 2
Legend
Patient 1
Patient 2
Patient 3-9
Positive: Leakage (Same Patient Same Class)
Valid: Possible Doppelgangers (Same Class Different Patient)
Negative: Can't be Doppelgangers (Different Classes)
b
Batch Correction sva::ComBat
Calculate PPCC between samples of different datasets
Group Sample Pairs By Similarity of Patient / Class
Calculate Cut Off Maximum PPCC of Negative sample pairs (Ignoring outliers)
Valid Sample Pair + PPCC > Cut Off = Doppelganger

Drug Discovery Today

Figure 2. PPCC principle

It is proved that PPCC data doppelgangers act as functional doppelgangers. One way to avoid the doppelganger effects is to put all doppelgangers in the training dataset or test dataset. By doing so, the accuracy is reduced to 0.5, which is the expected accuracy of the model trained based on random selection. Another way is to ignore doppelgangers. However, these two methods are not general and cannot suit for most cases. Wang mentioned three recommendations: (i) use meta - data as a guide to perform careful cross-checking. (ii) stratify data. (iii) perform extremely robust independent validation checks.

- **Adversarial Validation**

This method is used to detect data doppelgangers. To be specific, we build a classifier to try to predict which data rows come from the training dataset and which from the test dataset. If there are data doppelgangers, the classifier will be not able to distinguish them successfully. We can repeat the process for several times by eliminating one feature every time untill the classifier can learn to distinguish the training dataset and test dataset.

- **Dividing the model into Submodels**

By dividing the model into submodels, each responsible for one feature of the target subjects, such as metabonomics. Divide the training set and test set as usual, if a submodel performs well after the cross - validation and the accuracy is close to 100%, then the feature is likely to be a doppelganger. When establishing the classification model, only detected feature needs to be removed or less weighted, the possibility of doppelgangers are reduced (but not eliminated).

IV. Summary

In conclusion, the doppelganger effect is very common in the process of training and testing models, not only in biomedical data, but in all aspects. the doppelganger effect has an impact on the validation results of the model, leading people mistakenly think that the model is correct. There are some ways to reduce the doppelganger effect, but scientists still have a long way to go to eliminate it.

References

[1] Wang LR, Wong L, Goh WWB. (2021). *How doppelganger effects in biomedical data confound machine learning.* Drug Discovery Today.

[2] Alexander R, ISabelle S. (2011). *Data Doppelgänger: Addressing the Darker Side of Digital Identity.* European Advances in Consumer Research Volume 9, eds., 510.

[3] Robinson, SJ. (2008). *The doppelganger effect: spaces, traces and databases and the multiples of cyberspace.*

[4] Doll, K. (2022). *The Doppelganger Effect in Big Data, Explained.* Derived from https://www.shortform.com/blog/doppelganger-effect/.

[5] Orwant, J. (1994). *Heterogeneous learning in the Doppelgänger user modeling system.* User Model User-Adap Inter 4, 107 - 130.

[6] Rathgeb C, Fischer D, Drozdowski P, Busch C. (2022). *Reliable detection of doppelgängers based on deep face representations.* IET Biometrics Volume 11, 529.

[7] Jost, Z.(2020). *Adversarial Validation*. Derived from https://blog.zakjost.com/post/adversarial_validation/.