Doppelganger effect

Yixin He

In the press "How doppelganger effects in biomedical data confound machine learning", the author first described the definition of data doppelgangers: they are similar data that driven independently, when they occur in training and validation sets, they might lead the researchers misjudge the accuracy of some ML models. For example, if a model is trained and validated with highly similar data sets, it might perform better than it really is (called observed doppelganger effect). Chromatin interaction prediction system,

The author takes protein function prediction as an example, ML did a great job in predict the function of proteins with similar structure, but unable to predict proteins with similar function but different structures. Another example takes place in drug discovery, the quantitative structure-activity relationship (QSAR) models tend to consider structurally similar molecules with similar function.

A renal cell carcinoma benchmark data set is used by the author to explain doppelganger effect. To understand doppelganger effect, we first need to learn how to separate those doppelganger data from normal data sets. To do that, the author describes several methods that approach to identify data doppelgangers. Traditional methods such as scatterplot is not applicable since data doppelgangers are not that distinctive in reduced-dimensional space. DupChecker use MD5 fingerprints to define whether samples are replicated. However, replicated samples are not data doppelgangers, instead, they are leakage issues. Another method mentioned by the author is called pairwise Pearson's correlation coefficient (PPCC), it distinguishes data doppelgangers from data sets when the value of PPCC is high for specific pairs of data, though it still shorts in measure the ability for data doppelgangers to influence ML missions, it is meaningful to use PPCC to identify data doppelgangers. The author used renal cell carcinoma (RCC) proteomics to construct the benchmark scenarios. Firstly, based on similarities of each sample pair's patient and class, sample pairs are sorted in 3 types: 1) positive: same patient and class; 2) valid: same class but different patient; 3) negative: different classes. Data doppelgangers cannot happen in negative cases since they are from different classes, or in positive cases since they are considered as leakage, therefore, high proportion of high value PPCC data doppelgangers are founded in valid cases.

After identifying PPCC doppelgangers, the author wants to check if they act like functional doppelgangers, witch will inflates ML's performances. The result is that PPCC doppelgangers do have inflationary effects. The magnitude of inflationary effect to MLs depends on models been chosen. For instance, k-nearest neighbor (kNN) models and naïve bayes models show clearer proportional relationship between the number of doppelganger pairs and the models' performance than decision tree models and logistic regression models. The author noticed that when all doppelgangers are putting together in the training set, accuracy of ML models is about 0.5 as expected,

indicates the elimination of doppelganger effect. Unfortunately, this is not a good way to avoid doppelganger effect in ML models, since the size of training set is fixed, including more PPCC doppelganger data means possibility of lacking knowledge, which influence ML model's function.

To avoid doppelganger effects, one way is to classify data sets more detailed. For example, the author split each chromosome and cell type to generate training-validation sets. Though it does helps with doppelganger effects, it needs prior knowledge as backup, and for small sample size, we cannot divide samples to more detailed groups. So far the author haven't come up with method to solve doppelganger effects that will not lead to large reduction on sample size or require large amount of prior knowledge backup. However, the author give 3 recommendations for mitigating doppelganger effects: 1) using meta-date as guidelines to perform cross-check, 2) classifying data, 3) validating ML models by many independent data sets.

For me, I don't think doppelganger effects are unique to biomedical data. Doppelganger data are randomly chosen during training and validating processes, it can happen everywhere in areas including application of ML such as big data and statistics. The problem is, they are hard to identify form normal data. To avoid doppelganger effect, we first need to find those doppelganger data, one way is to measure the PPCC value. Then, for large amount of data sets, there are 2 possible ways to mitigate doppelganger effect: remove those valid data doppelgangers founded by measuring PPCC value, or sperate the raw data into more detailed classes. For small amount of data set, we cannot remove data doppelgangers because it will cause models lack in knowledge, we can instead choose those models that influenced slighter by data doppelgangers.

References

L.R. Wang et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.017