Report on *How doppelgänger effects in biomedical data confound machine learning*

Abstract

Machine learning (ML) has become an advanced tool for medical drug discovery. However, the similarities between data pairs in public data set can lead to overestimation of ML model performance. The report summarizes the content in the paper and talk about my understanding of data Doppelgängers and Doppelgänger effects of other data types and the possible methods to avoid Doppelgänger effects.

Introduction

Machine learning methods can help scientists understand the mapping relationship between certain molecular structures and biological activity. Quantitative structure-activity relationship (QSAR) is an effective model for biomedical scientists. It represents molecules in binary sequences and use those sequences as input to a machine leaning model to study the contributions of different substituents at different positions on a molecular structure to a certain biomedical function[3]. However, recent studies indicate that a large quantity of the data in public data set are duplicated and can cause issues in validate a machine learning model. This is called doppelgänger effect. Several methods such as batch correction, calculating Pearson's Correlation Coefficient, outliner detection, are used to avoid the doppelgänger effects[4]. However, those methods fail to explain the relationship between data doppelgängers and doppelgänger effects and might not be practical in small dataset. The paper analyses the properties of data doppelgängers and doppelgänger effects in several ML models and the disadvantages of previous methods of mitigating doppelgänger effects, and finally proposes advices and future research in avoiding doppelgänger effects.

Data doppelgängers, doppelgängers effect and functional doppelgängers

Data doppelgängers is data pairs where two data are highly similar. Doppelgänger effect is the model performs well because of the existence of data doppelgänger. Data doppelgängers might not produce doppelgängers effect. When they do, they are called functional doppelgängers.

The paper points out that data doppelgängers is abundant in biological data. Take protein for example, proteins from the same ancestor have similar sequences and similar functions. If we use proteins from the same ancestor as train set and validation set, the trained model might fail to predict proteins that has similar functions but less similar sequences. if data doppelgängers exist in train set and validation set, the training results can't differentiate good-trained models from bad-trained model.

Identification of data doppelgängers and functional doppelgängers

The paper proposes a method to identify data doppelgängers. Firstly, calculate the pairwise Pearson's correlation coefficient (PPCC) between a pair of data. Then, group the data into Positive, Valid and Negative where Positive data pairs represent data leakage, Valid data pairs might be Doppelgangers and Negative data pairs can't be Doppelgangers. The data in Valid pairs whose PPCC is larger than the maximum PPCC of Negative pairs are identified as Doppelgangers.

The paper then analyses the effect of data doppelgängers and proves that they inflate ML model performance. Furthermore, the more data doppelgängers in train and validation set the more inflated the ML performance[1]. The paper also studies the inflation of different ML model performance with different dosage of data doppelgängers with perfect leakage and binomial as control group. The results show that different ML models' performance change differently with the increase of data doppelgängers dosages.

The paper also mentions methods to identify functional doppelgängers from general data doppelgängers. We might find a subset of validation set whose performance is good regardless of the ML model. That subset of validation set will be potential functional doppelgängers[1].

Methods to mitigate doppelgänger effects

The paper illustrates and analyses the disadvantages of several methods of avoiding doppelgänger effects. The first one is putting all the data doppelgängers in train set or test set but also points out that this method will lead to lack of information in fixed train set or the performance on validation set either good or bad. The second method is to develop a rigorous assessment system however it needs prior knowledge and good quality benchmarking data[1]. The third method proposes removing all PPCC data doppelgängers. These methods may not work for small data set with high percentage of PPCC data doppelgängers.

Despite the tricky nature of doppelgängers, the author provides several guides to guard against doppelgänger effects. First, using meta data to construct negative and positive samples to specify the range in which doppelgängers exist. Then, put all the doppelgängers into train set or validation set. Second, split the validation set into different groups and test the model on different groups. Third, use strictly independent validation set like testing the model in real life scenarios.

I think one strategy to avoid doppelgänger effect is to find the atom data pattern that match the function. We can gradually increase the resolution of annotations in data set to test if the pattern is atom pattern.

Another strategy is to use synthetic data as validation set[2]. [2] combines the chemical groups randomly and produce the end-point label based on the accumulated scores of the chemical patterns it has which contribute to multiple functions.

I propose a method which might evaluate the proportions of data doppelgängers in test set by calculating the distance between test set data and their counterparts (have counter labels). If the distances are small, then the test set is a bad test set because it's very similar to train set, if the distances are large, the test set is better and might have less data doppelgängers. First, train the ML model on train set. Second, add a changeable modification Δx on each test set data x, Δx could be zero, and use the data $x + \Delta x$ as input to the trained model, find Δx that maximize the loss function of the model without changing the model weights. Finally, evaluate those Δx to see if they are big enough to produce a good test set. Fig.1 shows the illustration of the method.

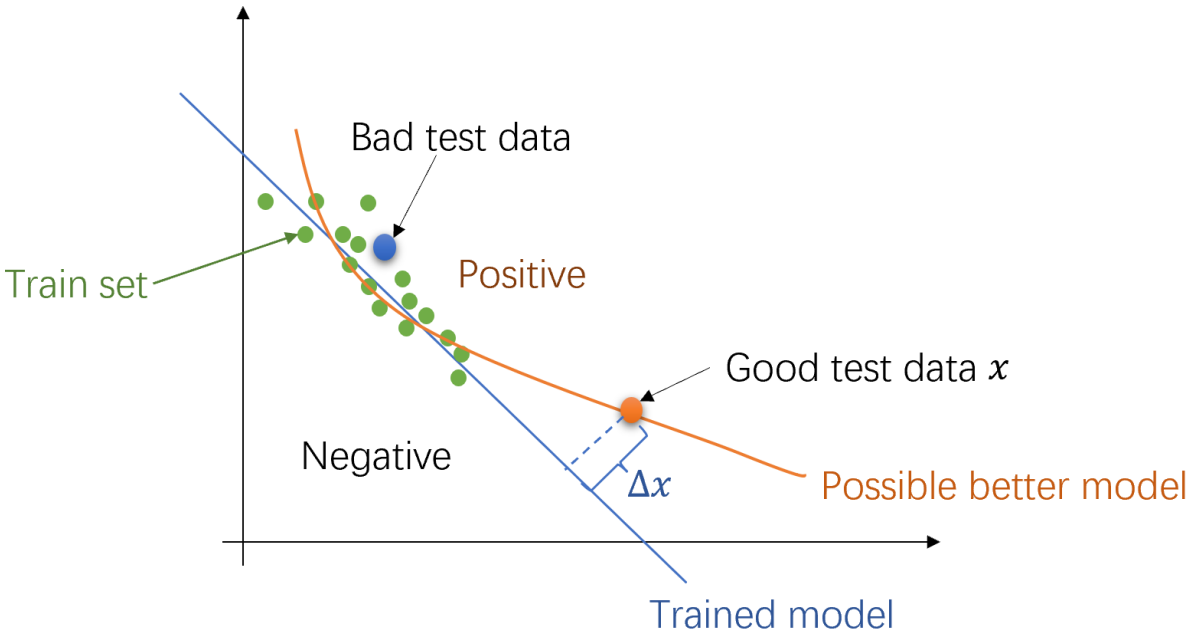

Figure 1 Illustration of the method

Are doppelgängers effects unique to biomedical data?

In my opinion, doppelgänger effects are not unique to biomedical data. I think doppelgänger effects exist in other data type as well. The reason why biomedical data have doppelgänger effects is probably because the whole dataset distribution is unknown and there is correlation relationship among sequences of genes, proteins and other biochemicals coming from the same data source. Other data type also might have those properties. For example, in the industrial defect detection of cables, defects like raptures have different morphologies, but the data set acquired by camera have the same kind of rupture patterns even the cables are from different sources. It's because that those cable photos are captured under the same environment circumstances. However, different cables may be in different environments, for example, some cables are in a wet environment, such as being immersed in water, while others are in a drier environment. Ruptures formulated in different environments might have different patterns. The ML model which only use photos captured from cables under same circumstances as train and test set might suffer from overestimation of model performance. The ruptures on cables from different sources form data doppelgänger. Because of the existence of doppelgänger pairs, the machine learning model fail to detect cables ruptures in other environments and produce doppelgänger effects.

Conclusion

Only when the validation data is independent of the training data, ML validation method is effective. However, this assumption is usually considered correct without prior examination, doppelgägers is quite common in the test data, it will directly affect ML accuracy. Two main factors will affect the extent of this inflationary effect, namely the similarity of functional doppelgängers and the proportion of functional doppelgängers in the validation set. To avoid performance inflation, it is important to check whether there are potential doppelgängers in data before assortment in training and validation data.

Reference

[1] Wang, Li Rong, Limsoon Wong, and Wilson Wen Bin Goh. "How doppelgänger effects in biomedical data confound machine learning." *Drug discovery today* (2021).

[2] Matveieva, Mariia, and Pavel Polishchuk. "Benchmarks for interpretation of QSAR models." *Journal of cheminformatics* 13.1 (2021): 1-20.

[3] Shoombuatong, Watshara, et al. "Towards the revival of interpretable QSAR models." Advances in QSAR modeling. Springer, Cham, 2017. 3-55.

[4] Waldron, Levi, et al. "The doppelgänger effect: hidden duplicates in databases of transcriptome profiles." JNCI: Journal of the National Cancer Institute 108.11 (2016).