Doppelganger Effect

Machine learning models have been increasingly develop and cross-validation techniques are commonly used to evaluate these models. However, the reliability of such validation methods can be affected by the presence of data doppelganger. Data doppelganger means training data and validation data are highly similar because of chance or otherwise. And doppelganger effect means a classifier falsely performs well because of the presence of data doppelganger. Data doppelganger may not all guarantee a doppelganger effect. Data doppelganger that also generate the doppelganger effect are termed functional doppelganger.

In my opinion, doppelganger effects are not unique to biomedical data, they will also appear in other field, but it is probably more common in biomedical data. For example, in natural language processing, the statement structure in the training set and the validation set might be similar, making the classifier more effective on the validation set when performing the classification task. There are similar examples in the field of computer vision. For example, in the task of pedestrian re-recognition, if everyone dresses the same in the verification set as he does in the training set, and everyone dresses differently from each other, it is easy for the model to complete this task. However, in the application, people's clothes cannot be invariable, and there is a high probability of people wearing the same or similar clothes. In this case, the model may have obvious errors.

But we should be more concern about the doppelganger effect in biomedical data, because a lot of time biomedical data might be used to drug research, which is a matter of life. For example, in the case of drug discovery, people sort similar molecules with similar activities into both training and validation sets which will confound model validation, because poorly trained models might still perform well on these molecules. People can only differentiate poorly trained models from their well-trained counterparts by testing their performance on similar molecules with different activities.

In order to solve doppelganger effect, it is important to study whether data effects exist in training sets and validation sets. The method used today is mainly Pearson's correlation coefficient (PPCC). This method can be used to discover the relationship between the training set and the verification set. A very high PPCC value reflects the data doppelganger in both sets. The disadvantage is that this approach cannot make a link between data doppelganger and their ability to confound ML tasks. So can only identify whether there is data doppelganger, not whether there is a doppelganger effect.

Directly removing data doppelganger and its potential doppelganger effect is difficult to achieve today, and we can only do as much as possible to avoid the effect. One approach is to put the identified doppelganger data with potential doppelganger effect into the training set or verification set as much as possible through the PPCC value, so as to prevent duality effect and allow a more objective evaluation of ML performance. Another approach is to divide the data into hierarchies with different similarities, such as PPCC data doppelgangers and non-PPCC data doppelgangers. Through this method, we can evaluate the performance of different levels of the model. In the poor performance level, we can understand the weakness of the classifier and point out the direction of further improvement. In addition, we can also carry out independence test to clarify the objectivity of the classifier.

In addition, we can take some ideas from other fields to ensure model robustness. For example, many times in order to judge model performance, we will use lots of datasets to compare and judge. For biomedical data, there may be few similar fields but completely different data. But if there is, we can try to analyze the performance of the model through multiple datasets, so that we can ensure that the model is robust. We can also expand the coverage of our data by collecting data from different regions or different times, and to some extent reduce the probability that the training set and the validation set are highly similar.