Identify and avoid doppelgänger effects during machine learning

The use of machine learning (ML) increased the efficiency of drug discovery and development process. However, the classification model based on ML sometime resulted in doppelgänger effects that confounded the accuracy of classifier and ML. This paper showed how doppelgänger effects are prevalence in biomedical data and provided evidence that ML can be confounded by doppelgänger effects. Ultimately, the author gave recommendations to resolve the issue caused by doppelgänger effects, that is to identify data doppelgänger before training-validation split.

Doppelgänger effect occurs when training and validation data sets are highly similar with each other so the classifier will perform well regardless of the quality of training. Data doppelgänger exists in various aspects of biological data. For instance, for certain validation data, even their features were randomly selected and were far from a validated data, a good performance can still be guaranteed given a particular training data. Another example was the protein function prediction in bioinformatics. If the model was established based on the assumption of similar protein sequence leads to similar protein function, then this model would be unable to correctly predict protein functions for proteins with less similar sequence but similar functions. Therefore, good models can be identified by testing their performance on similar objects (i.e., molecules or proteins) with different functions, given that these models are trained on informative structural properties.

Doppelgänger effect not only exists in biomedical data, but it also exists in all kinds of data generated from our daily life. For example, a person's data doppelgänger can be made up through lots of ways such as browsing history, status updates, GPS searching/routing history or

even credit card transactions. And this personal data doppelgänger is often utilized for personalized advertising: using certain computational models to predict which commodity you need most. Most of the time, this strategy is helpful as the commodities in your recommendation list are exactly what you are interested. But sometimes, the recommendation can be “creepy” probably due to Doppelgänger effect. For example, the computational models were trained to give recommendation based on customer’s username, age, IP address, and gender. And there happen to be two person who share the same age, IP address and gender but different usernames. Based on the training, the models predict those two people are the same person despite they have different usernames and give the same recommendation to both. Therefore, those two people would receive inappropriate recommendations resulted from doppelgänger effect.

To avoid Doppelgänger effect, we must identify data doppelgänger between training and validation data before validation. Earlier studies proposed to use the dupChekcer and the pairwise Pearson’s correlation coefficient (PPCC) to identify data doppelgänger but none of them work properly to constitute true data doppelgänger.