Report

In the *Machine Learning* course, this was the first time I divided the sample into three separate parts: training set, verification set, and test set. The training set is used to train the model, the validation set is used to determine the network structure or the parameters that control the complexity of the model, and the test set is used to test the performance of the optimal model. At that time, I discussed with my tutor whether there is a model that falsely performs well due to data similarity between the training set and the validation set. However, my tutor told me if the subsequent data is similar to the data in the validation set, it does not prove that the model does not work well. If the difference is large, then the model cannot be used. However, in this paper, I learn the concept of the doppelgänger effect and how this problem hinders the practice and development of biomedical data science.

Overview

This research is about the learning of the paper "How the doppelgänger effects in biomedical data confound machine learning", and I put forward some of my own opinions in the end.

Abundance of data doppelgängers in biological data

Data doppelgängers exist in many present fields of bioinformatics: chromatin interaction, protein function prediction, drug discovery, and so on. For instance, the performance of existing chromatin interaction prediction systems has been overstated because test sets share a high degree of similarity to training sets. In another example of the protein function prediction, based on the principle that proteins with similar sequences are presumed to be similar in function, some proteins with less similar sequences but more similar functions were falsely predicted. On the one hand, the principle related to the doppelgängers effect should get more attention, because it may play an important role in the whole experiment. On the other hand, if a model or principle is true in most cases (cases of data doppelgängers), can we consider using some other methods to deal with those special data falsely predicted?

Identification of data doppelgängers

There are several approaches to identifying the presence of data doppelgängers, and the most feasible one is the pairwise Pearson's correlation coefficient (PPCC), which could capture relations between sample pairs of different data sets.

The original PPCC paper never conclusively made a link between PPCC data doppelgängers and their ability to confound ML tasks, and the leakage is the reason for the doppelgängers effects. To identify the practical use of PPCC, the author used the renal cell carcinoma (RCC) proteomics data to construct benchmark scenarios.

When the author identified PPCC data doppelgängers based on the PPCC distribution of the valid scenario against the negative and positive scenarios, it seemed that the result was not sensitive enough.

Besides, the author checked PPCC distributions of the same tissue pairs, different tissue pairs, and replicates from the same tissue, finding that PPCC distributions are assuredly lower if we compare different tissue pairs, and PPCC distributions are extremely high for the same tissue pairs and replicates from the same tissue. That evaluations suggest that PPCC has discrimination value and can be used to identify PPCC data doppelgängers in RCC.

Confounding effects of PPCC data doppelgängers

The author explored PPCC data doppelgängers effects on validation accuracy across different randomly trained classifiers. The result demonstrates the presence of PPCC data doppelgängers in both training and validation data inflates ML performance, which is not due to the randomly selected features. The same situation may occur in every ML model. Moreover, the more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance.

The result of the experiment is that PPCC data doppelgänger produces inflationary effects similar to data leakage. K-nearest Neighbor (kNN) models and the Naive Bayes model show an evident similarity between eight doppelgängers and perfect leakage. The k-nearest neighbor models and Naive Bayes model in which both the training-validation set with eight doppelgängers in validation almost showed an identical accuracy distribution to the training-validation set with perfect leakage. However, different models are affected differently. kNN and Naive Bayes models have a clearer linear relationship.

A method to reduce the accuracy to what is expected is to constrain the training set or the validation set. However, it is not an optimal solution.

Ameliorating data doppelgängers

In the given paper, there are two methods mentioned. The first one from Cao and Fullwood is a more comprehensive and rigorous assessment strategy, based on the particular context of the data being analyzed. The second one is to use the PPCC outlier detection package, doppelgangR. However, the former needs a thorough knowledge of cells contextual/benchmarking data, and the latter is not applicable to small-size data sets.

Recommendations

To avoid doppelgänger effects, the author gives us three pieces of advice. Firstly,

perform carefully cross-checks using meta-data as a guide, which helps to identify potential doppelgängers and has a more objective evaluation of ML performance. Secondly, perform data stratification. Stratifying data into strata of different similarities may use different methods to solve different problems. Thirdly, perform extremely robust independent validation checks involving as many data sets as possible, which means it is more likely to inform on the objectivity of the classifier.

My point of view

We live in the era of data when machine learning is becoming more and more popular, and it exists almost everywhere in our daily life. In addition to biomedical data, other fields also have doppelgänger effects. For instance, in the used-car markets, machine learning is used to estimate the price of a used car, thus giving people who want to sell a car advice about the specific time of selling a car to get the highest yield. Used-car data contains many variables, such as car price, registration time, display mileage, transmission type, number of transfers, car brand, number of accidents, and maintenance accidents. These data, combined with the current situation and trend of the second-hand car market, would indicate the price of each used car. The data is divided into a train set and a validation set. Used cars with similar characters are inferred to have almost the sample price. However, sometimes the model is unable to correctly predict the price of the used car with less similar characters but actually sold a similar price to a certain type of car. That's when doppelgänger effects happen.

There are some measures we can take to avoid doppelgänger effects in the practice and development of machine learning models for health and medical science. Firstly, a more detailed and organized database is needed. In the area of data, data has become more precious. Many databases are not open to the public. A data platform to share information is needed. And there should be strict standards for data classification and data entry. If there were such a detailed classification and accurate data platform, perhaps more data doppelgängers could be discovered before validation.

Secondly, maybe we could make some changes in training and validation sets. Using the enzymes that are dissimilar in sequence overall but with similar active site residues (mentioned by the author) as an example, when we meet this kind of "special" data, maybe we could put it into training sets from validation sets and put the rest of data in the database into new training sets and validation sets. This may make the model have more different data, which may mitigate the doppelgänger effect. Besides, all the identified data doppelgängers from both training and validation sets are selected, some of which are put into the training sets (other normal data in the training sets is still there). After that, use some new data to build the validation sets. I suppose this may also lessen the doppelgänger effect.

Thirdly, the choice of machine learning method is very important. Every method has its own preference advantages and disadvantages. Maybe different data doppelgängers

in different fields of health and medical science use more fit models may mitigate the doppelgänger effect.

There is also a thing I want to mention. The first time I read that paper, I almost misunderstood the definition of overfitting and doppelgänger effects. But after I read it more times, I found that the reasons causing these two problems are different and the results are different too. doppelgänger effects are because of the similarity data in both the training and validation sets, and it makes the models falsely perform well. And overfitting is because the learning of noise and models truly performs badly. So I guess maybe there would be a relationship between them. When there are many similar data in training sets, the model may be overfitting at the same time with the data doppelgänger, and the model may perform better than only having the doppelgänger effects. That means overfitting may cause even more serious doppelgänger effects. So when we think of avoiding doppelgänger effects, we may need to guarantee it is not overfitting.