Report

In *Biomarker Screening and Its Mechanism Analysis Based on Machine Learning, Statistics and Bioinformatics*, after I screen microRNA based on the TCGA data set, my tutor emphasized that I should use the data from other source to ensure that these microRNAs are significant difference expression between tumor tissue and normal tissue. This is the first time that I construct awareness to validate results by using the different data set. But I know little about its potential problems. In this paper, I learn a concept of doppelganger effect and how this problem hinder the practical application and development of biomedical data science.

Overview

This research demonstrates an adverse effect, namely doppelganger effect, in machine learning models is inevitable and the potential method to mitigate it.

Abundance of data doppelgangers in biological data

Data doppelgangers is common in biomedical data. For example, the performance of chromatin interaction prediction systems has been overstated because of the similarity between test sets and training sets. It can also influence the prediction of protein functions. If we just consider the proteins of similar sequences indicate that they may have similar function, we will miss the situation that proteins with less similar functions have the similar structure. It means our understanding of biological principle is important, because it can play an essential role in our data analysis, whatever the analysis method we used. On the other hand, the dimensions of data are also important. In this case, it will be better if we consider both of the sequences and functions of proteins instead of just its sequences.

Identification of data doppelgangers

Although some methods are already tried in identification of data doppelgangers, they are not feasible and reliable enough.

The pairwise Pearson’s correlation coefficient (PPCC), captures relations between sample pairs of different data sets. It can’t make a link between PPCC data doppelgangers and their ability to confound ML tasks and constitute incorrect data doppelgangers because of the leakage, but it has potential to identify functional doppelgangers from PPCC data doppelgangers.

In order to identify the practical significant of PPCC, authors constructed benchmark scenarios by renal cell carcinoma (RCC). Authors conduct negative cases, valid cases and positive cases by using different type of data. They observed a high proportion of PPCC data doppelgangers on the valid scenario, but its distributions exist as a wide continuum, which means we can't set a specific cut-off to screen the data doppelgangers.

Then, authors checked PPCC distributions between same and different tissue pairs. For same tissue pairs, PPCC values remain high overall, because of the regulators which genes share. For different tissue pairs, PPCC values are lower. For the replicates from same sample or tissue, PPCC value are extremely high. Hence, PPCC has reliable discrimination ability.

Confounding effects of PPCC data doppelgangers

Subsequently, authors wanted to know whether the PPCC could influence the machine learning performances by using different randomly trained classifiers.

The results demonstrated that PPCC data doppelgangers in both training and validation data inflates ML performance, but the result shouldn't be like this due to the randomly selected features. The same results emerged in each ML models. In conclusion, the more doppelganger pairs represented in both training and validation sets, the more inflated the ML performance. There is obvious dosage-based relationship between the number of PPCC data doppelgangers and the overstatement of ML performance.

Thus, PPCC data doppelgangers can confounds ML outcomes. K-nearest neighbor (kNN) models showed the most similarity distribution between eight doppelgangers and perfect leakage. kNN and naïve bayes models have a clearer linear relationship between performance inflation and doppelganger dosage than decision tree and logistic regression model.

Placing all doppelgangers in the training set will eliminate the doppelganger effect. But constraining the PPCC data doppelgangers to either the training or validation set doesn't the optimal solution. In the former, PPCC data doppelgangers will occupy the limited spaces when the size of training set is fixed and models can't learn well. In the latter, the doppelgangers will all either be predicted correctly or wrongly.

Ameliorating data doppelgangers

To ameliorate data doppelgangers, Cao and Fullwood tried to create a particular context of data by splitting training and test data based on individual chromosomes, but it's hard to implement.

Using doppelgangR to remove PPCC data doppelgangers also has adverse effect, because the removal of PPCC data doppelgängers would reduce the data to an unusable size.

Recommendations

Authors illustrates 3 methods to guard against doppelganger effects. Firstly, performing cross-check using meta-data, which allow us to set a feasible score range to assess doppelgangers and assort them all into either training or validation sets. Secondly, we can stratify data into strata of different similarities and evaluate model performance on each stratum separately. Thirdly, we can perform robust independent validation checks involving as many as possible.

In the future, we could explore other methods to directly identify the functional doppelgangers. For instance, we can find a subset which can predicted correctly in any ML methods and pairing this approach with PPCC to identify the similar part between training and validation sets.

My viewpoint

It is indisputable that we are living in the era of data, in which data is conducted and used all the time. In addition to biomedical data, other fields also face the doppelganger effects. For instance, banks need to use machine learning method to predict the probability of their clients to pay credit card in the future and its amount of money. Each client data contains many variables, such as amount of the given credit, gender, education, marital status, age and history of past payment. The data is divided into train set and test set. If some client's basic information is similar, especially for one type of credit card (some credit card has unique target user), doppelganger effects may happen and the reliability of the machine learning models may decrease.

There are some implements we can do to mitigate the doppelganger effect in health and medical science. Firstly, we have a lot of room to improve in data source. In China, although more and more researchers and doctors realize the importance of medical data, they haven't overcome the obstacles of the data source yet. For the sake of use ratio of data, we need to complete the information sharing platform by collecting more comprehensive data and uniform the standard of them. Especially the data collection, we need a standard to describe them precisely and enrich the feature of them. For example, when describing the data of tumor tissue and normal tissue, we should consider it as thorough as possible, by adding details like pathological classification, stage and therapeutic method. And then we may attain more reliable validation results.

Secondly, we may use some strategy in training and testing. We can divide data set into training and test sets repeatedly and randomly, means one data can participate in either of them every time. Besides, we can also make use of some statistical method, like t-test, SAM, volcano plot or ROC, as references to evaluate the validation results.

Thirdly, we may perform Data Augmentation to attribute more value on current data. Data Augmentation is widely used to avoid over fitting, especially when the dataset is small. We can process images by flipping, rotating, scaling, cropping, translating or adding gaussian noise to enrich our data amount and variety. This method may mitigate the doppelganger effect, although it doesn't hedge against data doppelgangers directly.

In addition, when performing the interaction among protein and other molecules, the protein we chose may share similar sequences and lead to over fitting when conduct models. To solve this problem, we can cluster the similar or same sequences in different class after and remain only one sequences in each class.