Report

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, for model itself, we can adjust its parameters for best performance, like kernel function, regularization parameter, learning ratio and iteration time. Suitable parameters are helpful to improve the performance.

I find that the choice of machine learning method is important too. We should consider following things before choosing the algorithm: the size of training sets, the dimension of data, whether the data are linearly separable, whether the features are independent, our tolerance of bias and variance, etc. Each algorithm has its strength and weakness, we can maximize the former one and minimize the later one. For example, Naive Bayes is easy and quick, but it always cause high bias. Support Vector Machine is well-performed in both linearly separable and inseparable data, but it means high complexity and storage-occupying. Random Forest's parameters are easy to be adjusted, but it may induce over fitting. Neutral Network is powerful, but it's difficult to set suitable layer, activation unit and activation function, and its training is really time-consuming.

Secondly, for data, in addition to enrich its diversity, I think our understanding of data is also essential. The value of data depend on our cognition, especially for biomedical data. For example, as the paper said, some protein share similar function, although their sequences are different. So we can't just predict protein function in the easy way like sequencing alignment. We should label more feature to each protein and analysis the multidimensional data in advanced algorithm.

Multi-omics integration is the area always attract my attention. We can explore many value in integrated omics data. But one of the most difficulties scientist facing is the understanding of complex omics data, which can't be compensate by well-performed machine learning algorithm. So we should enhance our cognition of biomedical issue and it depends on experiment and clinic deeply. Therefore, biomedical data science can not develop without traditional biology and medicine.

We can also detect other coefficient's potential to identify the doppelganger effect like PPCC. This paper demonstrates a feasible and rational way to achieve this.

Finally, We can also detect other coefficient's potential to identify the doppelganger effect like PPCC. This paper demonstrates a feasible and rational way to achieve this.

In conclusion, I understand better about data doppelgangers after learning this paper and relative books. I also construct my basic knowledge structure of machine learning algorithms and I realize its miracle. In my opinion, data science can open-up huge possibilities in this era of data, especially in biomedical area. It is a discipline which can solve real-world problems substantially.