1，Introduction

Doppelganger effect indicates that similar training and validation data will result in an inflation of model accuracies on the validation dataset. The Doppelganger effect leads to good performance regardless of the training set, which can exaggerate the role of machine learning in the application of the models. There are two forms of this effect, one is Data Doppelganger(DD) which refers to correlated or similar sample pairs that exhibit very high correlation. The other is Functional Doppelganger (FD) where we need to identify it before the models being trained.

2, Doppelganger effect in other areas

I think this phenomenon occurs not only in the biomedical field, but also in many other fields.

The phenomenon is not related to the domain of the data, but rather to the form and distribution of the data. I have found that the data on which this effect occurs has the following characteristics. First, using machine learning techniques and having separate learning and prediction sets. It also contents pairs of data. Second, it is possible that they are from different groups and have undergone some degree of pre-processing, leading to the existence of a clear division between the two indicators of sample and group. In the article (the one given) it can be found that the genetic data was chosen because it divides the negative positives well and is very clear. And there are not a few such data in other fields. For example, in the case of flight delays database, the data set includes delays record of planes from different airports and different airlines. We can see that the data can be divided into a variety of cases such as the same airport or the same airline, etc. The format of the data is very similar to that of the data set in biology where Doppelganger effect occur. I was finding that model fitting would start straight away after dividing the learning and test sets in most of cases, so the Doppelganger effect was rarely detected. After reading a number of papers, I found that after high model accuracy emerged, often researchers would take them for further discriminations rather than examining the nature of the dataset. In the field of bioinformatics, a relatively primitive and easy way to determine that a

machine learning model is out of calibration (i.e. there is a possibility of a Doppelganger effect) is to test the model on similar samples with different properties. However, when we analyse commercial or living data, we do not have as many samples with different properties to test, and tend to use principal component analysis combined with scatter plots to see the distribution of the samples, which also proves less likely to detect Doppelganger effect. This suggests that although not many are found, it does not mean that the species phenomenon exists only in the biological domain.

Another case in point is in text analysis. Text analysis is also similar to genetic data in biology, where we can think of articles as patients and words as influencing genes. Text analysis is multi-disciplinary and also employs a wealth of machine learning methods such as KNN, Parsimonious Bayes, Support Vector Machines, etc. However, I found that in a case study of text analysis for sentiment lexical analysis, the accuracy rate using multiple machine learning methods was as high as 99.5%, and each remained above 96%. Even with unimproved methods with obvious shortcomings, the accuracy of machine learning was very high. However, comparing the performance of the same set of data with non-machine learning methods, the machine learning algorithm did not change before and after the correction, nor did it change after replacing the database. I believe that there may be Doppelganger effect, which needs to be tested using more quantitative methods such as the PPCC data recognition procedure.

But at the same time, I also think that this phenomenon will be more frequent and easier to detect in biomedicine. Because biological data is more relevant when grouped and discussed by source, it is also easier to spot spuriously correct models.

3, methods to avoid the Doppelganger effect

The first thing we would want to do is to identify the Doppelganger effect, often using methods such as principal component analysis to reduce the dimensionality or drawing a scatter plot to observe it. But these types of methods are not very practical here.

First, metadata is a structured description of information and has the function of managing, examining and safeguarding data to achieve accuracy, consistency and specificity. After obtaining the metadata for weighting, we can use metadata for cross-checking. The metadata helps us to predict the range of PPCC scores and allows us to look for Doppelganger effect. This allows us to make a treatment of them, placing them in different test sets and training sets.

Second, data stratification is very common in the field of machine learning. We often perform data layering when the same data has different levels of meaning in different dimensions. Here we can also perform data layering. We divide the data into different layers. We should analyse the state of the model on each layer rather than evaluating the performance of the model on the entire test data.

Third, performing independent validation, where we can use as many validation methods as possible and look at the problem more objectively.

In which it makes sense to use machine learning methods to filter for Doppelganger effect in the R language. The R package called the "doppelangerIdentifier" is very useful to find Data Doppelganger, and it basically uses the first idea mentioned above. The first step is to use the metadata to find DDs in the data. Then, display the PPCC distribution of the sample pairs and make observations. The third step, testing, observing the accuracy of the data inflation. In the fourth step, visualisation. It can be used to help us determine if this is Doppelganger effect and if it is present next we can use a hierarchical approach for further modelling.[1]

4, citation

[1]. Wang LR, Choy XY, Goh WWB. Doppelgänger spotting in biomedical gene expression data. Iscience. 2022 Aug;25(8):104788. DOI: 10.1016/j.isci.2022.104788. PMID: 35992056; PMCID: PMC9382272.

[2]. Levi Waldron, Markus Riester, Marcel Ramos, Giovanni Parmigiani, Michael Birrer, The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles, JNCI: Journal of the National Cancer Institute, Volume 108, Issue 11, November 2016, djw146