I don’t think doppelganger effect is unique to biomedical data. It may also appear in other data type but in extreme rare case. So other data type doesn’t have inflation effect in testing model. For example:

It may appear in image data. There is a paper about building a video resolution classifier [1]. They used the scatter plot of payloads packets size against packets arrival time as data set which means each video traffic corresponding to a certain scatter plot. Doppelganger effects may happen in these scatter plots. For example, in 360p resolution video traffic data set, there may have two video traffic which captured from different video and in training set and validation set respectively but have a very similar scatter plot.

A picture containing schematic

Description automatically generated

Figure1

In health and medical science industry, doppelganger data has inflation effect which makes model more accurate than it should be when test it. For example, the kidney tumour cells of the patient in the paper were produced by mutations in normal cells at certain key gene nodes. Most of the gene fragments are identical in human cell. This makes for an extremely high degree of similarity between tumour cells and tumour cells and between normal cells and normal cells. The kidney cells are mutated at certain specific gene nodes that make them turn into kidney tumour cells, and some of the kidney cells are mutated in the same way, which allows different patients to have very similar kidney tumour cells. These are the doppelgangers.

As can be seen in Figure 2, the more cells at high PPCC values, the greater the number of doppelganger cells in the two samples (it also could be data leakage). PPCC was used to describe the linear relationship between the cells in the two data sets.

Chart, scatter chart

Description automatically generated

Figure2

Because of the similarity of the renal tumour cells, it is inevitable that when you randomly select renal tumour cells from different patients to test your model, it is possible that some of the selected cells are doppelganger with the renal tumour cells used to train the model. This results in your model testing looking much better than it should be. Based on Figure3, you can see that the cells used to test the model may form doppelganger cell pairs with multiple cells used to train the model. This makes the inflationary effect even greater. In detail, there are 18 cells in the figure, yet 26 pairs of doppelgangers are generated.

Diagram

Description automatically generated

Figure3

However, not in all cases the higher the number of doppelganger cell pairs, the higher the accuracy of the test of the model (the stronger the inflation effect). This is related to the choice of model. KNN as well as naive bayes models have a clear linear relationship between performance inflation and doppelganger dosage compared with decision tree and logistic regression models.

The doppelganger data is a frequent occurrence in biomedical data, so it is important to avoid or eliminate its effects in order to train a good model and test it successfully.

The first approach is to use all of the data to train the model, which seems to solve the doppelganger problem, but still results in inflation when looking for data to test the model. This is because there is likely to be a doppelganger of renal tumour cells in new patients in relation to the renal tumour cells used for training. This is because human cells have too many of the same genes and may mutate at the same genetic nodes. My guess is that if different species of kidney tumour cells were used perhaps the effect of doppelganger could be reduced.

A second approach would be to use the PPCC outlier detection package to detect doppelganger data, and then remove them to mitigate the doppelganger effect. However, this approach cannot be used when the training data is small and the proportion of doppelganger data is high, as this would leave too little data for the training model to be trained.

The third method is cross-checks using meta-data as a guide. With the information of meta-data, we can identify potential doppelgangers and assort them all into training set or testing set. The meta-data in this article refers to the PPCC score range from -1 to 1. The closer the result is to 1, the more linearly the data are positively correlated with each other.

The fourth method is to try not to use meta-data, but to find subsets of the data used to validate the model that can be correctly determined without using the machine learning method and then remove them. This is followed by testing with non-doppelganger data. This will increase the objectivity of the model test.

Reference:

[1] Kamal, Ali & Bukhari, Syed Muhammad Ammar Hassan & Khan, Muhammad Usman Shahid & Maqsood, Tahir & Fayyaz, Muhammad. (2022). Traffic Pattern Plot: Video Identification in Encrypted Network Traffic.