Data Doppelgänger Effect

LIU Jianheng

liujianheng@buaa.edu.cn

1. Introduction

The data doppelgänger effect is gradually gaining more attention in the context of the data explosion and the massive deployment of machine learning algorithms in many industries. Data doppelgänger is a common phenomenon in fields such as biomedicine, and its presence can negatively confound machine learning algorithmic models. Therefore, it is important to effectively identify data doppelgänger and design solutions with versatility to address doppelgänger effects in the corresponding systems.

2. Data Doppelgänger Effect

The word Doppelgänger comes from German, where doppel means double and gänger means walker. Usually, the doppelgänger effect refers to two people looking identical, or to an apparition in human form. This concept was introduced in the field of data science to describe the appearance of two data sets with very similar characteristics, and the phenomenon is called data doppelgänger. Typically, when evaluating the model performance of a machine learning based classifier model, both the training and test datasets should be derived independently from the full data[1]. When a pair of data doppelgängers are present in the training and test sets, this may have a negative confounding effect on the classifier model, resulting in an inflated performance in training and testing and a weaker performance in real-world scenarios. This phenomenon is known as the data doppelgänger phenomenon.

Data doppelgänger effects are widely observed in modern bioinformatics. For instance, machine learning algorithmic models are widely used in protein function prediction. In their study Goh and Wong observed that proteins with similar sequences caused the doppelgänger effect. Machine learning models inferred from similar sequence features that the corresponding proteins were descended from the same ancestral protein, further making the inference that they had similar functions, which did not always correspond to the actual situation.

The data doppelgänger effect is not only present in biomedical data but is also widespread in other fields and data types. In the field of imaging systems, for instance, visually assisted pre-diagnosis systems are of great importance and have received a great deal of attention in the treatment of diseases such as skin cancer. In my previews project experience, due to the different types of skin cancer and the difference between benign and malignant, it is possible that the visual characteristics of the affected area may be similar, but the actual disease may be different, which is a consequence of the doppelgänger effect. Apart from that, in the field of gene sequencing, when data show that two individuals have similar genetic sequences, it may be inferred that they have a similar risk of developing a particular disease, when this may not be the case. In addition, data doppelgänger effect may also occur in the domain of recommender systems, when two users have similar browsing records, the recommendation system may presume that the two users have similar interests and further recommend other recommendations from the other user, which may not be effective.

3. Quantitative Understanding of Data Doppelgänger Effect

From a quantitative point of view, the data doppelgänger effect arises mainly because the data corresponding to the data doppelgänger are close together in the sample feature space and appear in both the training and test sets. Since the data points in the test set have already influenced the model during training, the closer the points in the training set with the same label, the higher the probability that they will be correctly classified by the model, but in fact this sample distribution may not reflect the true data distribution, which can lead to negative confounding effects during the training of models such as classifiers.

In a real system, I suppose there could be several common causes of this situation. The first cause may come from sampling, that is, the sample distribution of the collected dataset differs significantly from the true distribution. In the case of completely random sampling, it may be helpful to increase the amount of data collected from the sample. In the case of non-completely random sampling, it is important to consider what factors contribute to such a distribution bias, which may include factors such as age, geographical location, etc. The second contributing factor may be that the features chosen are not good enough to train a good classifier, at least not at the sample space where data doppelgänger appears. This is a good time to consider adjusting the introduced features or performing specific feature engineering, which I will discuss in more detail in the next section. A third factor may be the choice of model, which by its very nature may not be as robust to different data types and data distributions. Work such as data visualisation can help to select a more suitable model to mitigate the data doppelgänger effect.

4. Checking for Data Doppelgänger Effect

The pairwise Pearson's correlation coefficient (PPCC) is a data doppelgänger identification method[1]. The unusually high PPCC values indicate the presence of data doppelgänger. However, In the experiments, the PPCA values may still be high even when the same samples are considered, probably because PPCA possesses a possible meaniful discrimination value, while in practice it may be difficult to constitute a quantitative relationship between PPCA and the confounding effect of machine learning models.

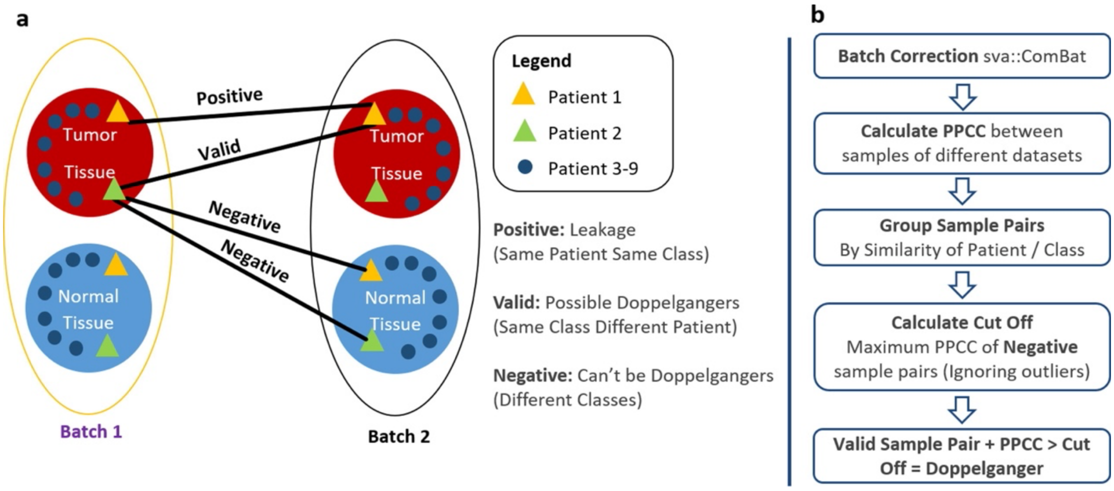
a
Positive
Valid
Negative
Negative
Tumor Tissue
Normal Tissue
Tumor Tissue
Normal Tissue
Batch 1
Batch 2
Legend
▲ Patient 1
▲ Patient 2
● Patient 3-9
Positive: Leakage (Same Patient Same Class)
Valid: Possible Doppelgangers (Same Class Different Patient)
Negative: Can't be Doppelgangers (Different Classes)
b
Batch Correction sva::ComBat
Calculate PPCC between samples of different datasets
Group Sample Pairs By Similarity of Patient / Class
Calculate Cut Off Maximum PPCC of Negative sample pairs (Ignoring outliers)
Valid Sample Pair + PPCC > Cut Off = Doppelganger

Figure 1 Diagram of the pairwise Pearson's correlation coefficient (PPCC)

In addition, Wang also suggest performing careful cross-checks using meta-data as a guide, data stratification, and extremely robust independent validation checks as data doppelgänger is complex in real datasets[1].

5. Avoiding Data Doppelgänger Effect

Based on the quantitative understanding in the last sector, there are several solutions to avoid Data Doppelgänger Effect:

a. Feature Selection:

By introducing other features, or even additional data sets, it helps to bring the sample distribution closer to the true distribution and to mitigate the negative interference effects of the data doppelgänger effect. This is a relatively simple and operational approach, but in fields such as biomedicine there are generally problems of difficult data availability and uneven sample distribution, therefore a specific analysis of the sample is required.

b. Feature Extraction:

The sample space can be altered by feature extraction, and furthermore the distribution can be altered to improve the dispersion of the sample distribution, especially in areas where data doppelgänger occurs. Typical methods include sequential backward selection, genetic algorithms, particle swarm optimization (PSO), LASSO. Apart from that, there are automatic methods to generate new features such as principal components analysis (PCA), kernel tricks, and autoencoders.

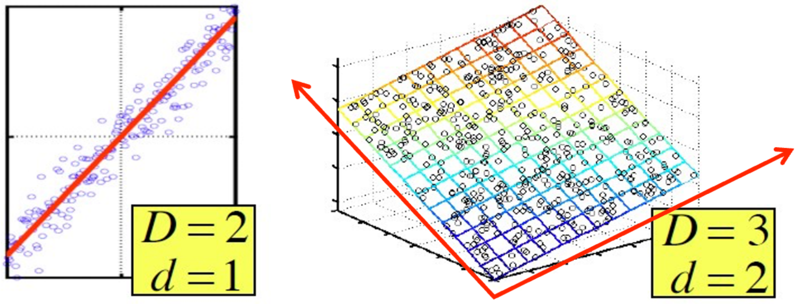

Figure 2 An Example of PCA in Feature Engineering

c. Model Selection:

By using ensemble learning methods, which allows weighted combination of multiple classifiers, the robustness of the whole classifiers can be enhanced. Typical methods such as AdaBoost and random forest are less likely to be overfitting and are also easy to train under complex data conditions.

6. Summary

In conclusion, doppelgänger effect is widespread in many domains and can have a negative confounding effect on machine learning classifiers. Due to their complexity, checking and avoiding data doppelgänger effect are challenging in the field of data science. Yet by solutions of feature selection, feature extraction, and model selection, it is still possible to attenuate the negative effects of doppelgänger effects.

Reference:

[1] Wang, L. R., Wong, L., & Goh, W. W. B. (2021). How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*.

[2] Goh, W. W. B., & Wong, L. (2019). Turning straw into gold: building robustness into gene signature inference. *Drug discovery today*, *24*(1), 31-36.