NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Doppelgänger Effects in Machine Learning

C2327982

Yuwei Jiang

jyw20001025@gmail.com

1 Introduction

With the development of digitalization, machine learning is one of the most important branches of artificial intelligence. The main task of this technology is to guide computers to learn from data. In machine learning, algorithms are continuously trained to discover patterns and correlations from large data sets, and then make optimal decisions and predictions based on the results of data analysis[1]. These algorithms can adaptively improve performance as the number of samples available for learning increases.

Machine learning is now found in a wide range of industries, including industrial production, biomedicine, economics and finance, and agriculture. Face recognition, tumor detection, stock analysis, and other machine learning applications have literally changed our lives. In the field of biomedicine, machine learning models are being used for drug discovery. The ability of machine learning to quickly search for and locate targets has greatly improved the efficiency of drug discovery and testing, and reduced R&D costs. Machine learning has also provided solutions for the treatment of COVID-19. However, the evaluation of the effectiveness of a machine learning model is impacted by something called "data doppelgängers."

2 Data Doppelgängers and Data Doppelgängers Effect

In the training and testing process of ML classification models, data doppelgängers occur when the training and validation data sets have high similarities, either by chance or otherwise[2]. And with the existence of data doppelgängers, the machine learning classifier can still yield good results regardless of the quality of training. Such false classification performance of a machine learning model caused by data doppelgängers is defined as an observed doppelgänger effect[2].

In the field of biomedical data science, doppelgänger data occurs when two have similar symptoms, features, or biomarkers. Such as protein function prediction gives us a false impression of high accuracy using doppelgängers data, but it performs for prediction proteins with less similar sequences but similar functions[2]. How to overcome the doppelgängers effect is now a great challenge for biomedical researchers.

While, from my point of view, the doppelganger effects can be widely seen on many other occasions, not just in the biomedical field. For example, during my independent research of the customer repurchase prediction, there are more than 40 million user behavior data obtained as my data set[3]. I try to transfer the problem as a classification task, label 1 for the customers who will repurchase, and label 0 for the customer who won't. It is generally believed that two customers with similar user behavior will most likely make the same purchase decision. And if I try to map one customer's behavior records to another who has similar ones, that could be called doppelgänger data. During the data pre-processing time, I occasionally found that there were data sets that were similar in their behavior records but had different final repurchase decisions. The classifier with too much doppelgängers data in the above case may result in poor performance of accuracy.

3 Data Doppelgängers Identification

Data doppelgänger effect can largely confound the prediction of ML models, so it is essential for us to detect the doppelgänger data before using it in the models. Using logical approaches like ordination methods or embedding methods combined with scatterplots to check the distribution of data would be unsuccessful because of its indistinguishable characteristics in reduced-dimensional space. Wang et al. [2] chose to use the methods of pairwise Pearson's correlation coefficient (PPCC) to capture the relations between sample pairs of different data sets. An anomalously high PPCC value indicates that a pair of samples exist as PPCC data doppelgängers. Although Wang recognized that the reported data don't constitute the true doppelgänger data, this is a theoretically possible method.

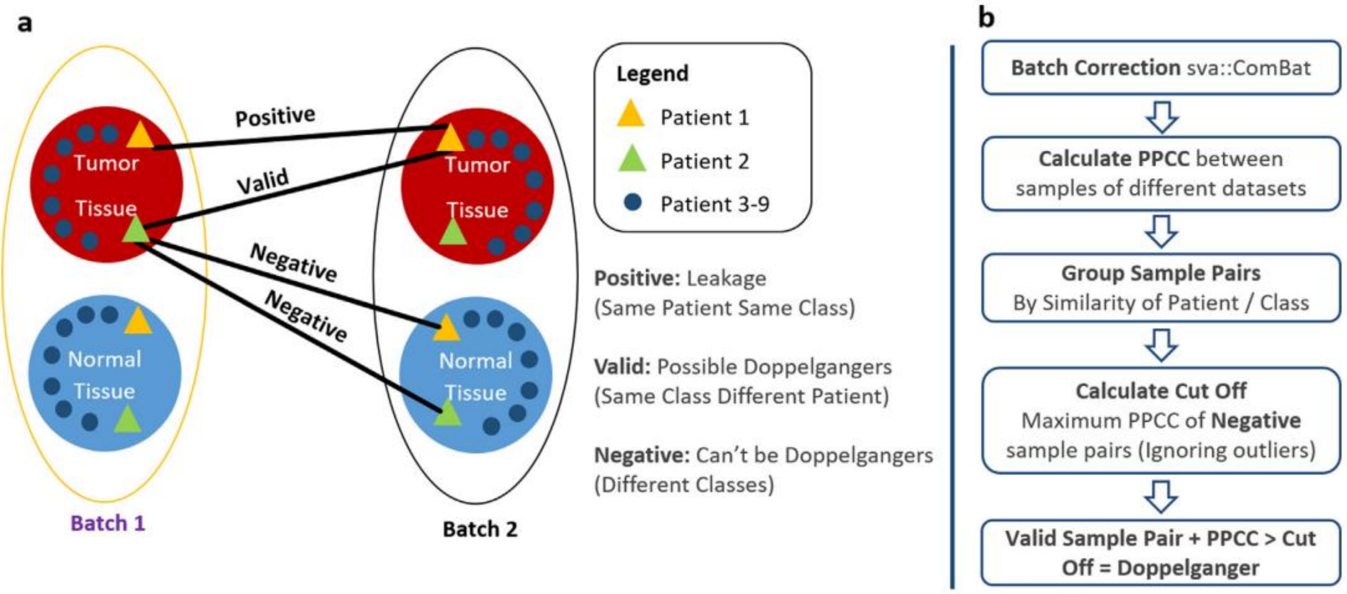
a
Positive
Valid
Negative
Negative
Tumor
Tissue
Tumor
Tissue
Normal
Tissue
Normal
Tissue
Batch 1
Batch 2
Legend
Patient 1
Patient 2
Patient 3-9
Positive: Leakage
(Same Patient Same Class)
Valid: Possible Doppelgangers
(Same Class Different Patient)
Negative: Can't be Doppelgangers
(Different Classes)
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. Pairwise Pearson's correlation coefficient (PPCC) data doppelgänger identification method

4 How to Avoid the Data Doppelgängers Effect

There are several ways to identify and avoid the data doppelgängers effect[2].

The First Method——Cross-Checks with meta-data

Meta-data will help us reconstruct the data set. We are able to identify probable doppelgängers using the meta-data and group them all into training or validation sets, effectively eliminating doppelgänger effects and enabling a more objective assessment of ML performance.

The Second Method——Data stratification

We can stratify test data into groups with varying degrees of similarity (such as PPCC data doppelgängers and non-PPCC data doppelgängers) and evaluate model performance on each group separately rather than on the entire test data set.

The Third Method——Robust independent validation check

Using extremely strong independent validation tests with as many datasets as possible (divergence validation). Several validation procedures can enhance the classifier's objectivity even though they are not a direct defense against data doppelgangers. Additionally, it shows how universal the model is.

5 Summary

In general, it might be difficult to deal with doppelganger effects while developing machine learning models for health and medical science. However, it is possible to lessen the effect of doppelganger effects and raise the precision and dependability of machine learning models in this field by carefully pre-processing the data, using methods like Cross-Checks with meta-data, Data stratification or Robust independent validation check mentioned in the article. Although these methods can reduce the doppelganger effects to some extent, there are still better solutions for us to explore in the future.

Reference

[1] Ji Shouling, Du Tianyu, Li Jinfeng, Shen Chao, Li Bo. A Review of Machine Learning Model Security and Privacy Research [J]. Journal of Software,2021,32(1): 41-67

[2] Li Rong Wang, Limsoon Wong, Wilson Wen Bin Goh, How doppelgänger effects in biomedical data confound machine learning, Drug Discovery Today, 2022, 678-685, ISSN 1359-6446

[3] Data set of Taobao users' shopping behavior, https://tianchi.aliyun.com/dataset/649