NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Data Doppelgänger Effect in Machine Learning

HU Xin Wen, Elisa
huxinwen23@163.com
Application Report for Msc Biomedical Data Science

1 Introduction

In this era of data explosion, with the development of machine learning technology, more and more traditional industries have been given a new light as a result. In recent years, machine learning models have been widely used in the fields of biomedical. Among other things, the use of machine learning models for modelling, analyzing, and predicting the effects and actions of drugs has improved their development effectiveness. As a result, it has become a popular tool for selecting instrumentalists prior to clinical trials. However, there exists data doppelgänger that effects the evaluation of the effectiveness of machine learning model.

2 Data Doppelgänger and Doppelgänger Effect

Doppelgänger effects refer to the phenomenon where two or more individuals have very similar or identical characteristics or features, such that they can be mistaken for each other or for a single individual. Wang *et al*[1] defined data doppelgänger that in machine learning models, a pair of data should be obtained independently are very similar. These similar data both allocate in the training set and the test set, causing the machine learning model preform surprisingly well in the training model, no matter how they trained. Hence, it will not preform as well as they trained when they are actually used in the reality testing. Such falsely classification performance of a machine learning model caused by data doppelgänger is defined as an observed doppelgänger effect[1]. Yet, Data doppelgänger may not guarantee a doppelgänger effect. So the data doppelgängers that also cause doppelgänger effects are termed fuctional doppelgängers[1]. In the context of biomedical data, doppelganger effects can occur when two or more individuals have similar medical histories, symptoms, or biomarkers, which can lead to confusion or incorrect decisions being made when analyzing the data.

Data doppelgänger not only occurs in the biomeidcal data. Hnece, it is possible for doppelgänger effects to occur in data from other domains, such as imaging, gene sequencing, and metabonomics. For example, in imaging data, two individuals may have similar imaging features, such as tumor size or shape, which can lead to confusion or incorrect diagnosis. In gene sequencing data, two individuals may have similar genetic variations, which can lead to incorrect predictions about their likelihood of developing certain diseases or responding to certain treatments. In metabonomics data, two individuals may have similar metabolic profiles, which can lead to incorrect predictions about their overall health or risk of developing certain diseases.

Doppelgänger effects can occur in text data, such as commentaries or reviews, if two or more pieces of text are very similar or identical. This can lead to confusion or incorrect decisions being made when analyzing the data, particularly if the text is being used to make predictions or classifications. In my final year project, which is an empirical analysis on investor sentiment and stock price based on text analysis. I crawled four hundred thousand data about one single stock from a online stock communication community, in order to

construct the investor sentiment index. During the process of data pre-processing, I discovery that people often use common stock market phrases or "proverb", which can lead to a lot of near-identical comments on the same trends of "up"and "down". Correspondingly, the frequencies of these phrases will rise. As a result, it caused doppelgänger effect when I extract the feature terms based on the TF-IDF.

3 Quantitative understanding

From a quantitative perspective, the doppelgänger effect refers to the phenomenon that a machine learning model trained on a dataset that is not representative of the target population, leads to poorer generalization and potentially biased forecasting when applied to a new situation.

For example, if a machine learning model is trained on a heavily male-biased dataset. As a result, when it makes predictions about females, it may make similar predictions about females as about males. This is regardless of the fact that the model has not been trained with enough data to accurately predict the likelihood of a woman developing a certain disease. In this case, the model may be subject to the doppelgänger effect because it makes similar predictions for women and men, even in spite of the fact that it has not been trained with enough data to accurately predict the likelihood of women developing the disease. If the model is trained on a dataset that is not representative of the target population, it may make inaccurate predictions when applied to individuals who are not well-represented in the training data.

Hnece, from a quantitative approach, the doppelgänger effect can be measured by comparing the model's performance on the training dataset with its performance on the hardened dataset or realistic data. If the model performs significantly less than training, this may indicate that the model is suffering from the doppelgänger effect.

Another way to quantify the doppelgänger effect is to look at the error rate or prediction accuracy of the model on different subgroups of the target population. If the error rate or accuracy of the model varies significantly between subgroups, this may indicate that the model is biased and does not generalise well to all members of the target population.

4 Avoid Data Doppelgänger

There are serval ways to avoid data doppelgänger:

1. Incorporating additional data sources and domain knowledge to fulfill the training data diversity. In my opinion, this method is especially important to biomedical data set. Because most of those having typical variables focusing on a specific disease. However, human body is full of mystery that only baseline variables cannot have well performance in model validation. By incorporating additional data sources and domain knowledge, such as using NLP techniques to analyze electronic health records, constructing casual inference index to add effective predictor. It can help to provide a more comprehensive view of an individual's health and may help to reduce the impact of doppelganger effects. On the other hand, if the target data contains information about genetic variations, it may be useful to incorporate knowledge about the functional consequences of these variations into the model.

2. Using ensemble models, which combine the predictions of multiple individual models, can be more robust and less susceptible to doppelganger effects than single models. This is because ensemble models can take into account the diversity of the individual models and can be less sensitive to the specific characteristics of a particular group of individuals. There are several types of ensemble models. including boosting, bagging and stacking models. For example, using random forest which is a type of ensemble model instead of decision tree. They are more resistant to overfitting than decision tree and the a final prediction is made by aggregating the predictions of the individual tree. Besides, they can handle high-dimensional and complex data more effectively than the decision trees.

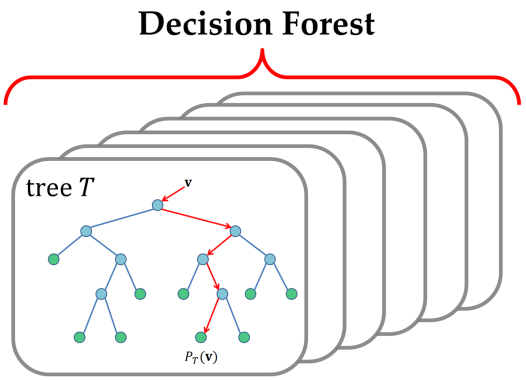

Figura 1: Random Forest

3. Using the Pairwise Pearson's correlation coefficient (PPCC) to capture linear relations between sample pair of different data sets[2]. An anomalously high PPCC value indicates that a pair of samples exists PPCC data doppelgänger. However, *Wang et al* considers that it never conclusively make a relationship between PPCC data doppelgänger and ability to confound ML tasks, so there is also a leakage phenomenon in the data using in this article[1],and the conclusion cannot prove the doppelgänger will expand the evaluation model. However, it's feasible in theory and worth to be referenced.

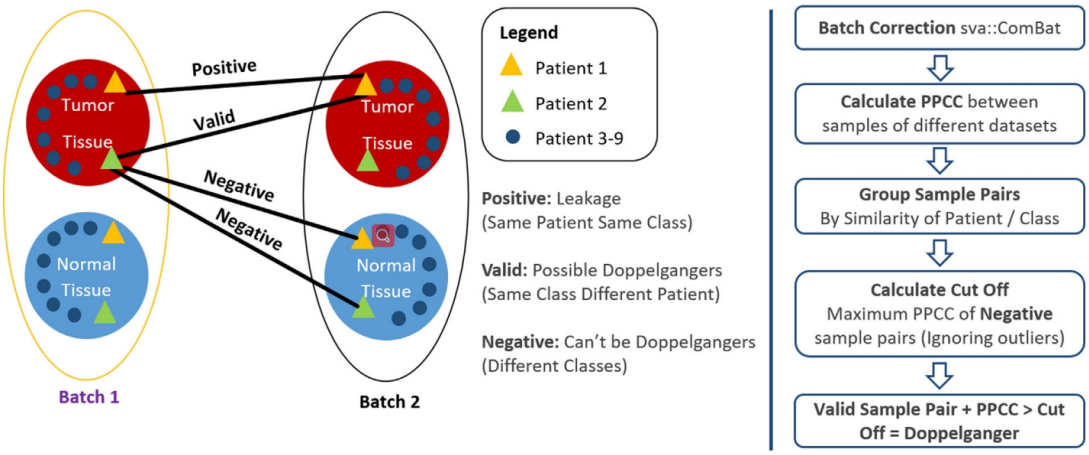

Figura 2: PPCC

5 Empirical Analysis

Based on chosen PPCC as the criterion to check for data doppelgänger by using the renal cell carcinoma (RCC) proteomics data *Wang et al*. RCC was chosen for its utility in constructing clear-cut scenarios, which has been divided in to three categories:

(i) Negative cases: Sample pairs are from different classes, in which doppelgänger are non-permissible.

(ii) Valid cases: Sample pairs are from different sample classes but assign with same label, but they may be doppelgänger. Its effect can be compared against positive cases.

(iii) Positive cases: Sample pairs from the same patients class and assign with the same label. These constitute obvious leakage issue and not considered are doppelgänger.

After divide the data into three categories, Group sample pairs by the similarity of class and eventually calculate the maximum PPCC of negative sample pairs. Eventually, identified PPCC data doppelgängers based on the PPCC distribution of the valid scenario is defined as valid sample pairs with PPCC value are greater than all negative sample pairs. From Figure 3(b), we can observed half of the samples are PPCC data doppelgängers with at least one other sample. Moreover, PPCC has meaningful discrimination value

which can be seen in Figue 3 (b). PPCC values o for some tissue pairs remain high, suggesting high correlations between samples, but the PPCC distribution are assuredly lower.

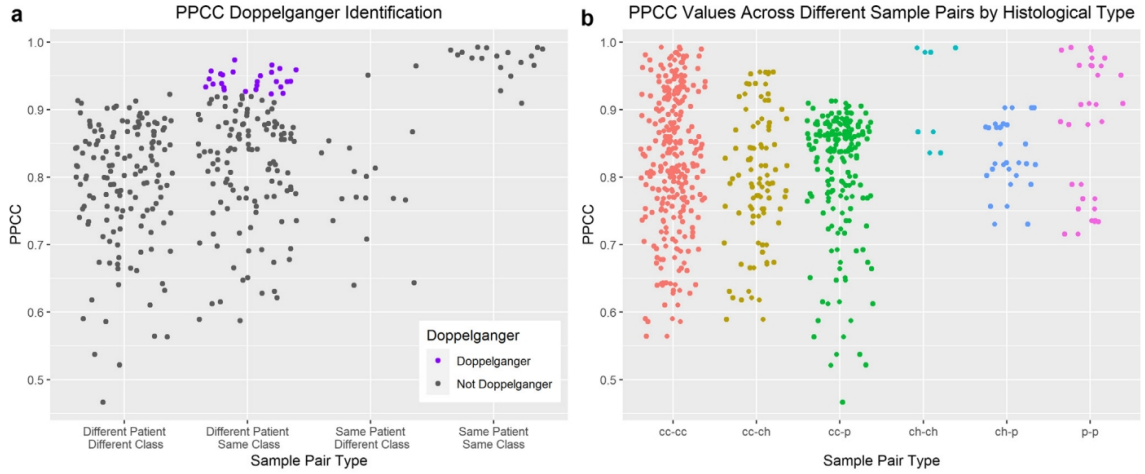

Figura 3:
(a) Distribution of pairwise Pearson's correlation coefficients (PPCCs) across different sample pairs.
(b) Distribution of PPCC values of different sample pairs by their histological types.

After data prepossessing and overview, *Wang et al* establish experiments through feature selection and train the data set with several machine learning models, including KNN, Naive Bayesian, Decision Tree and Logistic Regression.

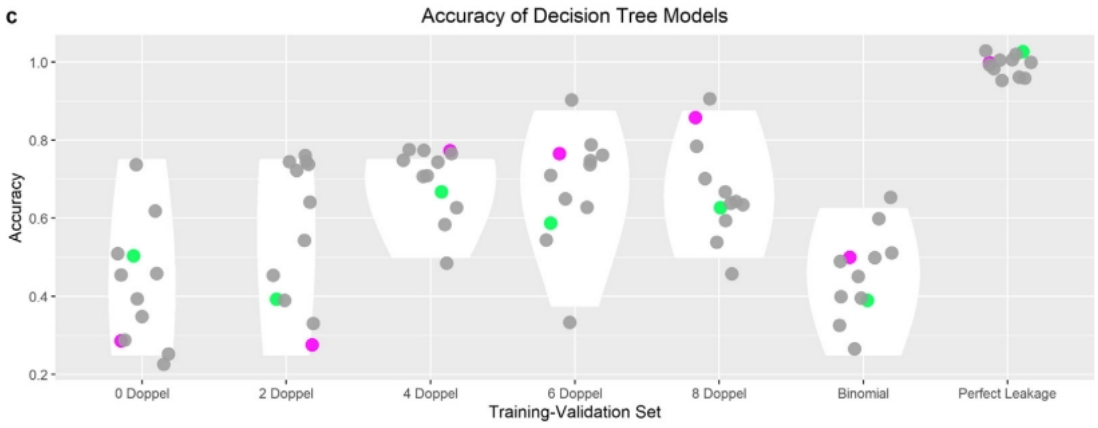

Figura 4: Result Eaxmple: Decision Tree

Example as the result of decision tree model, from figure 4, we cam observe that since the presence of doppelgängers in both training and validation data, has a strong effect on the model performance evaluation. Even the model of a meaningless random selection of features has a good training effect, and the more doppelgängers exist, the better the model performs and ML performance inflates. Through these experiments, *Wang et al* successfully proved that data doppelgängers will inflate the ML performance.

6 Summary

Overall, doppelganger effects can be a challenging issue to address in the practice and development of machine learning models for health and medical science. However, by carefully pre-processing the data, using machine learning techniques to identify groups of individuals with similar characteristics, and developing models that take into account multiple factors, it is possible to mitigate the impact of doppelganger effects and improve the accuracy and reliability of machine learning models in this domain.

Referências

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

[2] L.Waldron, M. Riester, M. Ramos, G. Parmigiani, M. Birrer. (2016). The Doppelgänger effect: hidden duplicated in databases of transcriptome profiles, *J Natle Cancer Inst*, 108.

[3] Beata Butryn, Iwona Chomiak-Orsa, Krzysztof Hauke, Maciej Pondel, Agnieszka Siennicka. (2021).Application of Machine Learning in medical data analysis illustrated with an example of association rules, *Procedia Computer Science*, 192, 3134-3143.

[4] Seong Ho Park, Kyunghwa Han. (2018). Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction, *Radiology*, 28(3).

[5] Jaber Alwidian, Bassam H. Hammo, Nadim Obeid. (2018), WCBA: Weighted classification based on association rules algorithm for breast cancer disease, *Applied Soft Computing, Elsevier*.