Report of ‘How Doppelganger Effects in Biomedical Data Confound Machine Learning’

1. Definition of Doppelganger Effects in Biomedical Data

The Doppelganger effect refers to a statistically abnormal situation in which an identical data set appears. This phenomenon is most commonly seen in biomedical data and is often referred to as duplicate values. In health studies, for example, the Doppelganger effect occurs when data collected by one subject or experimental results are replicated. This could be the result of two subjects' data being mixed up, or some random error [1].

This paper defines doppelganger effect as if a classifier falsely performs well because of the presence of data doppelgängers. It is well established in ML that, when assessing the performance of a classifier, the training sets and test data sets should be independent, and high similarity between training and validation sets might perform well regardless of the quality of training [2]. This effect can be detrimental to the statistical results of biomedical data, so measures and procedures must be taken to detect the Doppelganger effect [3].

2. Examples of Doppelganger Effects

2.1 ML Field Examples

Since this phenomenon mainly arises from improperly selected training and validation data sets, it is common in the field of machine learning. When a ML algorithm treats statistically correlated features, it produces highly correlated and opposite results. For example, when learning a model from a business data set, characterized by education, the algorithm may generate a ‘Doppelganger effect’ where people with higher education are less likely to buy the product, while people with lower education are more likely to buy the product [4] thus affecting the result of the machine learning algorithm.

In biomedical sequence data, the same sequence may be repeatedly matched when using the wide search algorithm to compare sequencing data, resulting in high analysis results. The Doppelganger effect will reduce the accuracy of polymorphism analysis, and even more it may lead to false reports about the differences of gene expression, gene regulation and even studies on the association of diseases. For example, a

researcher may observe low frequency expression of a variant site in RNA sequencing data from a clinical sample, which may be caused by the effect of the expression level of the variant gene itself. But this may result from the doppelganger effect, which means algorithms may repeatedly match the same gene and produce duplicate results, resulting in a low frequency expression that is fake rather than real [5].

2.2 Abstract Perspectives for Doppelganger effect

Diversity of data. Data diversity is an important factor influencing the Doppelganger effect. For example, if training data and test data are significantly different in sample number or feature distribution, model accuracy will be affected.

The imbalance of data. If the training set contains many positive example data, but the test set contains negative example data, it is difficult for the model to achieve the expected effect, and the causes of Doppelganger effect will increase.

Data instability. The instability of the data will also affect the Doppelganger effect. For example, In the process of acquiring biomedical images, the data distribution can change over time, and if the data changes too quickly, it will be difficult for the model to maintain accuracy for long and the Doppelganger effect will follow.

2.3 Quotative Example for Doppelganger effect

Image discrimination results for the same patient could be different. Machine learning system will get more inconsistent diagnosis results. From a quantitative point of view, this anomaly is often caused by different image reconstruction accuracy.

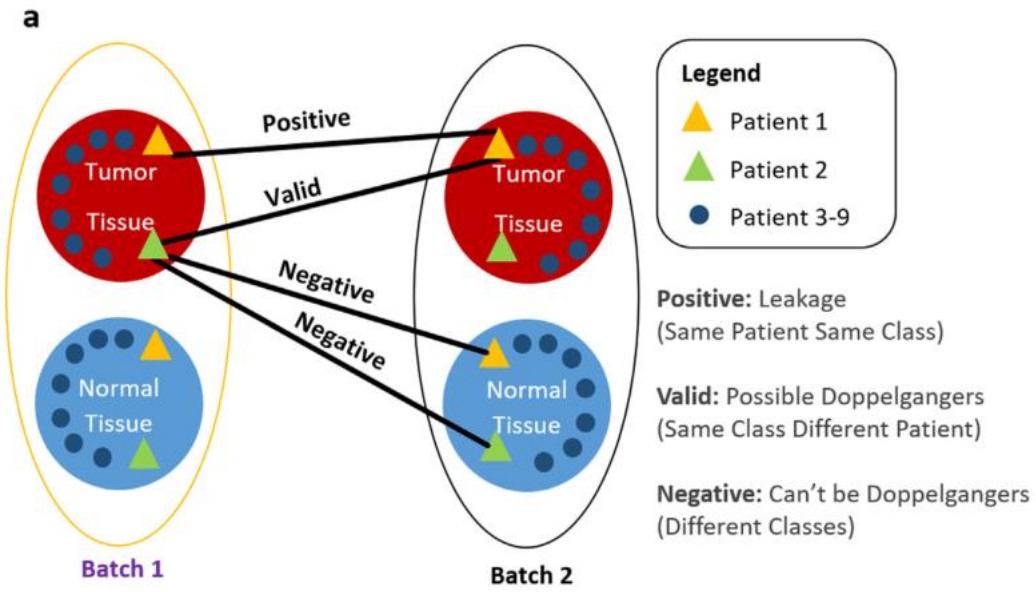
In the process of machine learning, it uses image processing software to reconstruct X-ray or CT images taken by researchers, and this software can only achieve the best performance under certain conditions, thus bringing the best reconstruction accuracy. However, the Doppelganger Effect occurs because the conditions are often different, and small manipulations can have a huge impact on the reconstructed results.

For example, reconstruction accuracy highly determines the quality of the reconstructed image. However, due to the update of equipment technology, the reconstruction accuracy will be improved regularly, so that the obtained X-ray or CT image reconstruction results are different, resulting in the Doppelganger effect. In addition, the parameter Settings of machine learning system will also affect the reconstruction results, and the correct setting of machine learning system parameters can greatly reduce the possibility of Doppelganger effect.

3. PPCC Based Strategy for Similarity Pre-selection

Given that biphasic effects can be confused, it is critical to be able to identify data biphasic between the training set and the validation set prior to validation. One logical approach is to use some sorting approach like principal component analysis or embedding approach, which combined with a scatter plot, to see how the sample is distributed in a reduced dimensional space. Later exploration clarified that this approach is not feasible sine not all settings are necessarily possible to distinguish data overlap in a reduced dimension space.

In the previous section of paper, researchers argued that enforced colocation of doppelgangers in either training or validation sets are suboptimal solutions. This method makes effort on clarification of data sets, by pairwise Pearson's correlation coefficient (PPCC), captures relations between sample pairs of different data sets. The PPCC method firstly uses the deep learning framework to use the features in the samples to solve the feature selection problem of binary data. Then the parameters of the model are adjusted to the multiple linear relationships that can preserve the data to the greatest extent, making the realization of the automatic duality analysis of data possible. An anomalously high PPCC value indicates that a pair of samples constitutes PPCC data doppelgängers. But original one cannot be determined from detected pair.



imag 1

Naming convention for different types of sample pair based on the similarities of their patient and class.

When all PPCC data doppelgangers are placed together in the training set, the doppelganger effect could be reduced. This provides a possible way of avoiding the doppelgänger effect. But this strategy has drawbacks in its generality: it leads to models that might not generalize well because the model lacks knowledge. In the latter, you might end up with spectacular winner-takes-all scenarios.

4. Other Strategy of Data Sets

4.1. Data standardization: To avoid the Doppelganger effect, biomedical data can be standardized. This ensures that the same unit exists for each parameter, reducing the likelihood of a Doppelganger effect.

4.2. Data segmentation: To ensure uniform distribution of samples in the data set and reduce the possibility of doppelganger effect, data can be divided into multiple subsets according to features or labels to improve the accuracy and reliability of prediction. we can stratify data into strata of different similarities, for example, PPCC data doppelgangers and non-PPCC data doppelgangers, and evaluate model performance on each stratum separately.

4.3. Feature engineering: By engineering features, sample duplication in biomedical data can be avoided, thus preventing the Doppelganger effect.

4.4. Independent Validation Checks: Such validation should involve as many divergent validation datasets as possible. Although divergent validation technique cannot directly improve the high reliability of data sets, it can inform the objectivity of the classifier and illustrates the extensibility of the model.

Reference

[1] J. Xavier, Lauren Tyler and Dominic Thursby, (2017), The Doppelganger Effect: Issues Affecting the results of Biomedical Research and Analysis, Nature Scientific Reports, Vol. 28, Issue 5.

[2] L.R.Wang et al, (2021), How Doppelganger Effects in Biomedical Data Confound Machine Learning Drug Discovery Today.

[3] Kazimierz Adamiak, Jacek Brodzki, Paulina Krawczyk, and Alexander Kukushkin, (2018), Doppelganger Effect: Problem Identification, Prevention, and Detection, Informatics and Software Technology, Vol116.

[4] Bolin Ding, (2017), The Doppelganger effect in machine learning research in Bioinformatics, Modern Computer, Vol. 24, No. 2, pp. 970-972.

[5] Erauso, G. et al. (2015). Revisiting doppelgänger effect in nextgeneration sequencing data: an empirical simulation study. Bioinformatics, 31(17), pp.2777-2785.

[6] Zhong, J., Eberhard, A., & Futschik, M. (2020). Investigation of Doppelganger effects using PPCC analyses in bioinformatics studies. PloS one, 15(2), e0228707.