Doppelgänger effects: concepts, prevalence, and solutions

Introduction

The field of machine learning (ML) has gained widespread popularity in recent years due to its ability to analyze large datasets and make predictions about new data. However, one common problem that occurs in machine learning models is the doppelgänger effect (DE), in which a model achieves high performance on a validation dataset, regardless of how it was trained [1]. The DE can lead to overestimation of a model's performance and potentially erroneous conclusions about its generalizability to new data.

The article titled “How doppelgänger effects in biomedical data confound machine learning” illustrated the definition of DE, prevalence of DE in biomedical data, identification of the DE using Pearson’s correlation coefficient (PPCC), exploration of the DE impacts on validation performance, and ways to ameliorate the DE. In this report, we will explore the concept of the DE, its impact on various fields, and possible solutions to address it.

Background

The advent of big data and ML has revolutionized biomedical research, including drug discovery, precision medicine, and biomedical material design [2-5]. ML algorithms can analyze large datasets to identify new drug targets, predict drug efficacy, and optimize the development of biomaterials for tissue engineering by analyzing cell behavior and interactions with different materials.

However, the DE is a common phenomenon where poorly trained models may perform well in validation sets, decreasing the generalizability and robustness of the models. **The DE is not unique to biomedical data and can occur in any field where machine learning models are used. In computer vision, for example, doppelganger effects**

can occur if the training data contains images of animals in a particular habitat, the model may perform well on a validation set that contains similar images but fail to generalize to new environments or types of animals. In natural language processing, doppelganger effects can occur when the training set and the validation set contains ambiguous words. Re-use of tissue specimens is widespread in clinical genomic studies, creating the DE in publicly available datasets: hidden duplicates that, if left undetected, can inflate statistical significance or apparent accuracy of genomic models when combining data from different studies. In this way, the DE is similar to overfitting, both of which can lead to failure in real-world applicability.

Wang et al. (2022) introduced two key definitions – data doppelgängers (DDs) and functional doppelgängers (FDs) [6]. DDs are sample pairs that exhibit very high mutual correlations or similarities. For example, we may use pairwise Pearson's correlation coefficient (PPCC) to identify DDs such that sample pairs with high PPCCs are also referred to as PPCC DDs. On the other hand, FDs are sample pairs that, when split across training and validation data, result in inflated ML performance, i.e., the ML will be accurate regardless of how it was trained (It can be assumed that such models have not truly "learnt").

The DE can have significant negative impacts, particularly in biomedical data, where the cost of errors can be very high. For example, a machine learning model that misclassifies a disease could lead to incorrect treatment decisions and poor health outcomes for patients. Therefore, it is essential to identify and address the DE to ensure the reliability and generalizability of machine learning models. We can avoid the DE by decreasing the effect of DDs and FDs, respectively.

Methodology

Accurately estimating model performance is a key strategy for mitigating the

doppelgänger effect. Cross-validation and external validation are two approaches that can help decrease the effect of FDs and provide a more precise estimate of the model's performance. Cross-validation is a technique that partitions the dataset into multiple subsets and evaluates the model's performance on each subset (Fig. 1) [7]. This approach helps estimate the probability distribution of the model's performance on new data, exploring various combinations of sample pairs and enabling more informed decisions about the model's suitability for real-world use. External validation involves validating the model's performance on an independent dataset. This dataset is less likely to have the same idiosyncrasies or noise as the training set, reducing the risk of the doppelgänger effect. External validation can provide a more reliable estimate of the model's performance and increase confidence in its generalizability.

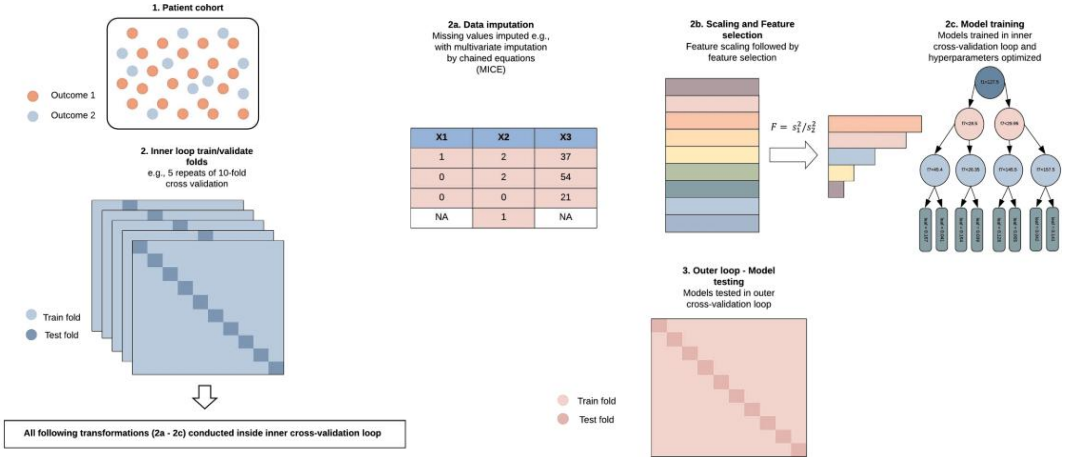

Fig. 1. Illustration to cross validation

Schnack et al. (2016) suggested that increasing the sample size can also help avoid overestimation of the doppelgänger effect [8]. Smaller samples tend to be more homogeneous, whereas larger samples are less susceptible to the effect of doppelgängers. By increasing the sample size, it becomes possible to introduce new pairs of similar individuals, reducing the proportion of similar pairs and enhancing the diversity and representativeness of the dataset. Consequently, models become more robust and better able to generalize to new data. Increasing the sample size is therefore a critical step in minimizing the doppelgänger effect and enhancing the accuracy and reliability of any analyses or predictions based on the data.

The DE arises from the "lazy" learning approach employed by many machine learning (ML) models, such as K-nearest neighbors (KNN), which seeks the most similar training samples to make predictions. To mitigate the DE, it is essential to improve the quality of the dataset by applying techniques like feature selection, feature engineering, or data augmentation. Additionally, constructing more robust models that can learn the underlying patterns instead of relying solely on sample similarity is another promising approach to decrease the DE.

Discussion

This report reviews the concept, prevalence, and possible solutions to doppelgänger effect. We have raised some important limitations of the solutions for addressing the doppelgänger effect. While the solutions discussed in the report are effective in addressing the doppelgänger effect, it is important to acknowledge their limitations and carefully evaluate the potential benefits and drawbacks based on the specific research question and characteristics of the dataset.

It is true that cross-validation may not be effective when there are a large number of similar pairs, as the similar pairs may end up in the same fold of the cross-validation and lead to overestimation of model performance. Therefore, it is important to carefully choose the number of folds and how the data is split in order to minimize the impact of similar pairs. Additionally, external validation is only effective when the external samples are heterogeneous and representative of the population of interest, which may not always be the case in biomedical research where the availability of external datasets may be limited.

Regarding the solution of increasing the sample size, it is important to consider the trade-off between increasing the sample size and potentially introducing more similar pairs. This is especially important in biomedical research where sample collection may be costly and time-consuming. It is important to carefully evaluate the benefits and

drawbacks of increasing the sample size based on the specific research question and characteristics of the dataset. Researchers should also consider alternative methods such as feature selection or dimensionality reduction techniques to increase the diversity of the dataset without increasing the sample size.

References

[1] Wang L R, Wong L, Goh W W B. How doppelgänger effects in biomedical data confound machine learning[J]. Drug Discovery Today, 2022, 27(3): 678-685.

[2] Chan H C S, Shan H, Dahoun T, et al. Advancing drug discovery via artificial intelligence[J]. Trends in pharmacological sciences, 2019, 40(8): 592-604.

[3] Singh A V, Rosenkranz D, Ansari M H D, et al. Artificial intelligence and machine learning empower advanced biomedical material design to toxicity prediction[J]. Advanced Intelligent Systems, 2020, 2(12): 2000084.

[4] Mishra R, Li B. The Application of Artificial Intelligence in the Genetic Study of Alzheimer's Disease. Aging Dis. 2020 Dec 1;11(6):1567-1584. doi: 10.14336/AD.2020.0312. PMID: 33269107; PMCID: PMC7673858.

[5] Levi Waldron, Markus Riester, Marcel Ramos, Giovanni Parmigiani, Michael Birrer, The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles, JNCI: Journal of the National Cancer Institute, Volume 108, Issue 11, November 2016, djw146, https://doi.org/10.1093/jnci/djw146

[6] Wang L R, Choy X Y, Goh W W B. Doppelgänger spotting in biomedical gene expression data[J]. Iscience, 2022, 25(8): 104788.

[7] Cearns, M., Hahn, T. & Baune, B.T. Recommendations and future directions for supervised machine learning in psychiatry. Transl Psychiatry 9, 271 (2019). https://doi.org/10.1038/s41398-019-0607-2

[8] Schnack H G, Kahn R S. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters[J]. Frontiers in psychiatry, 2016, 7: 50.