Introduction to Doppelganger Effects and Strategies to Avoid Them

OUYANG Hui

Introduction

While the techniques of machine learning have been used in various industries, there are still limitations and potential pitfalls to be resolved. In the article “How doppelgänger effects in biomedical data confound machine learning”, Li Rong Wang, Limsoon Wong and Wilson Wen Bin Goh stated the presence of data doppelgängers in the biomedical field and the accompanying impact on the performance of ML models. By their definition, data doppelgängers happens when independently derived data are very similar to each other. And because of the similarity between training and validation set, doppelgänger effect may occur and causing models to falsely perform well. Based on their arguments, this report mainly discusses that doppelganger effects are not only confined to biomedical data, and the possible solutions to avoid or reduce doppelganger effects and data doppelgangers.

Causes and examples of doppelganger effects

To explain if doppelganger effects occurs uniquely in the biomedical filed, how doppelganger effects arise and how it can happen in other fields should be clearly stated. The first reason doppelganger effects occurs is obvious, when training the ML model, the independence of validation set and training set is often assumed to validly evaluate the performance of the ML model. However, similarity can happen by chance or naturally. That is, it is possible to randomly pick data sets that have similar sample pairs and in the real world, certain categories will naturally produce more similar samples than others. As the independence and similarity of validation set and training set are usually unchecked, the doppelganger effects may be resulted. Moreover, this kind of problem would occur in any situation when applying the machine learning model, therefore doppelganger effects are not only occurring with biomedical data.

In addition, doppelganger effects can be resulted because of the nature of data itself. Although the independence of data is an essential precondition for cross-validation, dependence is pervasive in biological data. For example, doppelganger effects may emerge with quantitative structure–activity relationship (QSAR). While QSAR models assume that molecules that have similar structure will have similar activities, structure–activity pairs with high similarity in both training and testing sets will result in data doppelganger. However, it has been seen that this assumption is not always true. Sometime slight variations in molecule structures may produce

different biological activities, hence when predicating the activities of unusual new molecules, models trained with data doppelgangers will have poor performance.

It is worth noting that such similar feature–outcome relationships is not unique to QSAR model. Data doppelgangers can also be found when predicting protein–protein interactions with high correlations between pairs that share a same given protein. Other similar examples could be enhancer–promoter, regulator–gene and drug–protein interactions (Fig.1). Ignoring the dependence of data can cause the performance of models to be inflated (Whalen, Schreiber, Noble, Pollard, 2022).

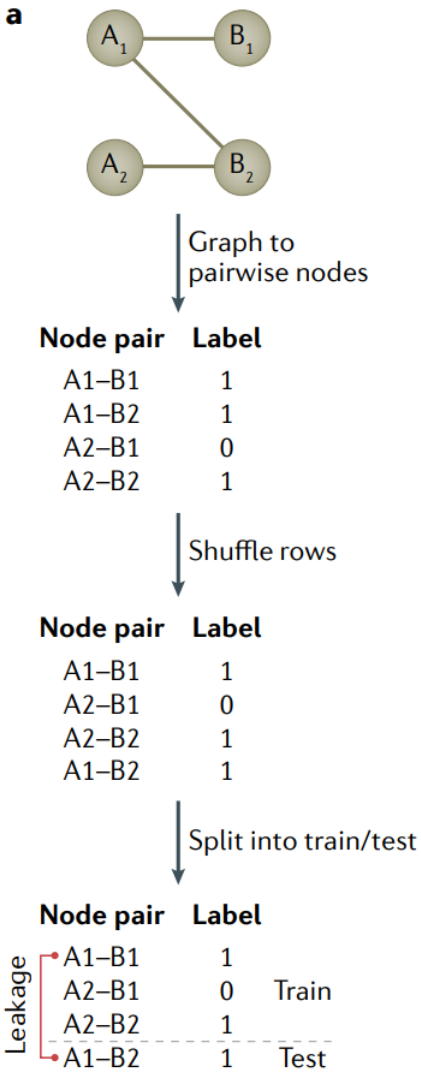

Fig.1. A graph shows how data can be dependent with each other, where sets A and B could be proteins and ligands, drugs and target genes or enhancers and promoters.

Strategies to mitigate doppelganger effects

With the understanding of why doppelganger effects, several solutions to eliminate or minimize doppelganger effects are proposed. First of all, it is suggested that before training a model, training and testing set should not have duplicate or highly similar samples. However, techniques to detect the samples that potentially cause doppelganger effects need to be improved, for example, the ordination method to distinguish data doppelgängers in reduced-dimensional space may only work in certain situation, and dupChecker can only identify duplicate samples rather than true data doppelgängers. Significantly, Wang, Wong and Goh (2021) illustrated that pairwise Pearson's correlation coefficient (PPCC) may be a better solution for detecting data doppelgängers and doppelgänger effects. Moreover, functional doppelgängers are detected if models can always have excellent performance when predicting subsets of validation data, and such subsets should be removed.

With the viable methods to identify potential functional doppelgängers, doppelgänger effects can be avoid by putting all data doppelgängers together in either training or data set. But the drawbacks of this approach are obvious: models will have poor generalizability. Another solution could be remove potential functional doppelgängers from the data set. Unfortunately, this only works with large amount of diverse data. When there are too many similar data in a small data set, this method will become unusable as remaining available data are not enough unless more advanced methods are introduced to train models with small datasets and can still obtain stable and trustworthy performance.

A more robust strategy to mitigate doppelgänger effects is to split training and test data based on the properties of the data being analyzed. In the case when dependence is naturally exist in the data sets, it is suggest that data in training or test set should from different groups. For example, when predicting the correlated functions of proteins, data collecting from the same family or complex should not place into training and test set at the same time (Fig.2). But such strategy requires a thorough knowledge of the related field (Whalen, Schreiber, Noble, Pollard, 2022).

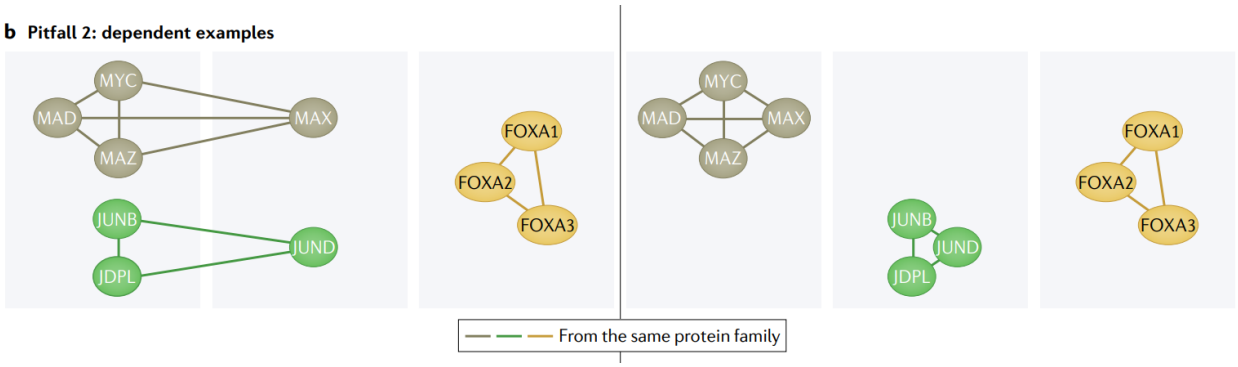

Fig.2. Left: Using entities in the same group. Right: Partitioning entire biological groups into either the training or test set.

However, approaches of manually splitting the training, validation, and testing sets can reduce the data doppelgängers but it does not guarantee to eliminate doppelgänger effects. Therefore, ensuring the models are trained with datasets with greater diversity and larger size is recommended. Other solutions may include developing the explainable model for biomedical field. As the aim of applying ML is to address real-world problems, the model should eventually perform well with real-world data, rather than just performing well in the validation set. Therefore to practically improve the algorithms and customize AI model for biomedical field or even for every single dataset would also be necessary for reduce doppelgänger effects and enhance the robustness and trustworthiness of the model.

Conclusion

In conclusion, this report explains that doppelganger effects are not unique to biomedical data by explaining that how doppelganger effects is caused during the data preprocessing stage and highlighting examples to show that some data types will naturally trigger doppelganger effects. Then possible strategies to mitigate doppelganger effects are discussed. Given that doppelganger effects will inflate the ML performance and cause the predictions to be untrustworthy, more efforts need to be put into improving methods to identify data doppelganger and developing ML model for biological field.

References

L.R. Wang et al. How doppelgänger effects in biomedical data confound machine learning. Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.017

Whalen, S., Schreiber, J., Noble, W.S. *et al.* Navigating the pitfalls of applying machine learning in genomics. *Nat Rev Genet* **23,** 169–181 (2022). https://doi.org/10.1038/s41576-021-00434-9