The Doppelgänger Effects on Machine Learning

Yingdong Zhang

1. Introduction

Nowadays, machine learning is increasingly being used in biomedicine. In the process, similar chromosomes[1] (Cao and Fullwood, 2019), RNA families[2] (Szikszai *et al.*, 2022), shared ancestry[3] (Greener *et al.*, 2022) and other independent studies have found that machine learning has an over-inflated verification performance between the training and verification sets. Wang et al., in How doppelgänger effects in biomedical data confound machine learning stated that the doppelgänger effect (DE) refers to the phenomenon that a classifier incorrectly performs well because of the presence of doppelgängers in the data[4]. Biomedical data has the characteristics of multi-dimension, massive and multi-source heterogeneous, so DE is very common in the biomedical field.

In the process of using machine learning to solve problems, we often evaluate experimental results by testing the accuracy of the model on the verification set. However, this method of evaluation loses its usefulness in the presence of DE. In addition, DE problems are common in test data and have a direct impact on the accuracy of machine learning. Therefore, the problem of DE should arouse our full attention.

This report will discuss the non-uniqueness of DE in data sets, analyze DE problems from a quantitative perspective, and make some opinions and suggestions on how to detect and avoid DE problems.

2. DE is Not Unique to Biomedical Data

DE is not unique to biomedical data, any data may have potential similarities or redundancy, so DE can occur in any type of data.

For example, I once worked on a project about traffic flow forecasting. In this kind of problem, if the data of the training and the test sets are both data from weekdays, then the changes in traffic flow are likely to be highly similar, and the model we trained can perform well in the test. However, when the model is applied in real life, especially when it is used to predict traffic flows during holidays (e.g., the Spring Festival), many problems may occur. Because during the Spring Festival, people do not need to work in the company but choose to go back to their hometown.

3. DE in Quantitative Perspective

Wang *et al.* quantitatively analyzed DE using the pairwise Pearson's correlation coefficient (PPCC) data[4]. PPCC is a measure of the linear relationship between two variables. It results in a value between -1 and 1, where -1 means completely negative correlation, 0 means no correlation, and 1 means completely positive correlation.

The researchers found that PPCC data in the training and validation sets improved the performance of the machine learning model even under the condition of randomly selected features. Furthermore, the more doppelgänger pairs represented in training and validation sets, the better the Machine Learning performance. When the "Doppel 4" data set was stratified into PPCC doppelgängers and non-PPCC data doppelgängers strata, all models showed better results on the PPCC data doppelgänger than on non-PPCC data doppelgänger. After experiments on several models, it is clear that PPCC data doppelgängers are related to the generation of DE. Especially on kNN and naive Bayes models, there is a clear linear relationship between the amount of PPCC data and the performance.[4].

4. How to Detect and Avoid DE

4.1. How to Detect DE

Wang *et al.* proposed the use of dimensionality reduction techniques for testing, that is, using ordination methods (e.g., principal component analysis) or embedding methods (e.g., t-SNE), coupled with scatterplots, to see how samples are distributed in reduced-dimensional space [4]. This approach can help identify similarities in high-dimensional data. However, for data that is indistinguishable in reduced dimensional space, this method is not effective. Besides, in DEs in Quantitative Perspective, we mentioned that PPCC data doppelgängers are related to the generation of DE, so we also can detect DE by measuring PPCC.

Researchers can use a data-layered approach to detect DE. They can divide the data into different, similar layers and then use the data separately to evaluate the performance of the model[4]. A lower-performing layer not only helps us find weaknesses in the model but also prevents DE from hiding those weaknesses and helps us detect DE.

DE can be detected by verifying the results. Researchers can try to verify the model with as many other data sets as possible before actually using the model. Even if DE exists in some data sets, as long as enough data is tested, the test results will converge to the true level of the model. Researchers can compare the results of the test set to the results of multiple data sets to detect the presence or absence of DE. In addition, some model evaluation methods (such as k-fold cross-validation) can be used to detect the presence of DE, and DE may occur if the model appears to be overfitted.

4.2. How to Avoid DE

Cao and Fullwood called for comprehensive and rigorous evaluation strategies based on the specific context of the data being analyzed to avoid DE[1]. However, this approach relies on prior knowledge and high-quality baseline data[4], so it is difficult to implement.

In Lakiotaki *et al.* 's study, PPCC data doppelgängers could be deleted to mitigate DE[5]. However, this method is only suitable for data sets with a low proportion of PPCC data doppelgängers. Because if the proportion is high or the data set is small, the data set will not be large enough for machine learning training after the PPCC data doppelgängers are removed[4].

Wang et al. proposed doppelgangerIdentifier, a software suite that eases doppelgänger identification in Doppelgänger spotting in biomedical gene expression data[6]. Researchers can use doppelgangerIdentifier as a tool to build training validation sets that will not overstate the accuracy of the model.

Before machine learning training, researchers should carefully select the features that are most relevant to the research task. If there are too many characteristics of data, there may be a lot of redundant data, and hidden duplicate information may appear in these data, thus leading to the occurrence of DE.

What's more, the researchers could try to optimize the algorithms for machine learning in the future, using deep learning or other techniques to further process potentially repetitive information in the data during training models. For example, ignoring information that is too similar in training.

5. Conclusion

In short, DE is a very important problem not only in the biomedical field but also in various fields. DE can decrease the performance of the model generated by Machine Learning in practice. We should fully analyze and process data before using Machine Learning to solve problems. After the generation of the model, DE should be further fully detected to reduce the influence of DE on experimental research. At the same time, we should also try to find new and effective ways to completely solve the problem caused by DE.

References

[1] Cao, F., & Fullwood, M. J. (2019). Inflated performance measures in enhancer-promoter interaction-prediction methods. *Nature genetics*, *51*(8), 1196-1198.

[2] Szikszai, M., Wise, M., Datta, A., Ward, M., & Mathews, D. H. (2022). Deep learning models for RNA secondary structure prediction (probably) do not generalize across families. *Bioinformatics*, *38*(16), 3892-3899.

[3] Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*(1), 40-55.

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

[5] Lakiotaki, K., Vorniotakis, N., Tsagris, M., Georgakopoulos, G., & Tsamardinos, I. (2018). BioDataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology. Database, 2018.

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