Along with the increasing application of machine learning (ML) in biomedical fields, researchers find it common while some models show perfect predictions on the validation set, failing in real-world applications. This stems from an interesting phenomenon called data doppelganger. Data doppelganger refers to independently derived data exist in training and test sets with similarity, it will dramatically inflate the performance of ML. This report will respectively illustrate the exclusivity issue of doppelganger effect in biomedical data and provide three effective methods to avoid it in the practice and development of machine learning models for health and medical science.

As far as its uniqueness is concerned, doppelganger effect is considered as non-exclusive to biomedical data. The foremost reason lying behind its widely use in multiple fields. It is generally acknowledged that doppelganger effect is not limited to biomedical data science. Additionally, until it is explicitly characterized, doppelganger effect tends to be confused with data leakage. A valid case is that the original paper of pairwise Pearson's correlation coefficient did not properly distinguish between them. Consequently, the concealment of doppelganger effect is believable when data scientists in various fields pre-process their data. For instance, Japanese scientists (Wu et al., 2019) have had good success with machine learning from a database of polymer materials. Doubts may exist in their dataset because they used a similar principle to quantitative structure- activity relationship (QSAR) (Chen et al., 2014) that similar structures have similar activity. They did not deliberately avoid data doppelganger, which would lead to over-fitting of doppelgangers’ features when performing machine learning on the training set. a

When abundant doppelgängers exist in the training and test sets, the loss function will decrease towards the direction of graduation after calculating into tensor, which would result in the Doppelgängers’ features being learned, following by inflation of models’ effects. In the experiment of this paper, all ML models showed higher performance on PPCC data doppelgängers than on non- PPCC data doppelgängers. This displays us how doppelgängers effect is generated from a quantitative aspect. And with doppelgangR Bioconductor package, doppelgängers can be detected in a convenient way. According to the mechanism of doppelgängers, three methods can avoid this effect to a large extent. Firstly, it is a good way to hold reserve an independent test set until the research is completed and use it for final validation test. If unable to keep a pure test set, collect some new data to make a new statistically pure test set then, or make the caveat that the new model was based on a choice made after observing the test set error. As a result, the performance estimate is likely to have an optimistic bias. Secondly, it is highly recommended to add noises and use regularized function to constrain parameters to make the model more generalized. The last suboptimal manner is to place doppelgängers in train sets or test sets. However, it has significant drawbacks: in the former, the samples in train sets will be monotonous while the model may not be able to distinguish all the doppelgängers in the latter. Consequently, a compromise approach may be superior to deleting pairs of doppelgängers with an extremely high value of PPCC and augmenting the remaining data by adding slightly modified copies of data which is not doppelgängers. In this case, lack of knowledge of short-flow learning can be avoided and doppelgänger effect can be adjusted.

|  |  |  |
| --- | --- | --- |
| **Models** | **Doppelgänger Accuracy** | **Non-Doppelgänger Accuracy** |
| K-Nearest Neighbours | 1 | 0.5 |
| Naïve Bayes | 1 | 0.5 |
| Decision Tree | 1 | 0.5 |
| Logistic Regression | 0.25 | 0 |

This report has respectively demonstrated the exclusivity issue of doppelganger effect in biomedical data and provide several effective methods to avoid it in the practice and development of machine learning models for health and medical science. More further explorations will be conducted in the future.

‌Wang, L.R., Wong, L. and Goh, W.W.B. (2021). How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*. [online] Available at: https://www.sciencedirect.com/science/article/pii/S1359644621004554?via%3Dihub#m0005 [Accessed 3 Feb. 2022].

‌

Wu, S., Kondo, Y., Kakimoto, M., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., Schick, C., Morikawa, J. and Yoshida, R. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Computational Materials*, [online] 5(1), pp.1–11. Available at: <https://www.nature.com/articles/s41524-019-0203-2>.

Q. Chen, L. Wu, W. Liu, L. Xing, X. Fan

**Enhanced QSAR model performance by integrating structural and gene expression information**

Molecules, 18 (2013), pp. 10789-107801