Report

Doppelgänger Effects

When the datasets are modeled with a large amount of highly similar data, the model may run the risk of performing falsely well, which is known as doppelgänger effects (Wang, Wong, & Goh, 2021). This is attributed to the appearance of highly similar data in both the train set and test set, thus making the model obtain the correct answer "directly" from the train set during test. This can result in the falsely high accuracy of the model when a test set is used with data doppelgängers. There is a typical example. The use of dataset to train the model can be regarded as an exam. The train set is equivalent to the questions that students have practiced before the exam, the test set is comprised of the questions encountered by students during the final exam, and the grade of final exam is treated as the model performance. If the final exam includes many questions the student encountered before, the student will achieve a high score. However, this high score may be worthless because the student has met the original question and known the answer, which is not truly reflective of the student's ability. Therefore, the excellent performance of a model may be untrue if the data in the test set is highly similar to the train set.

Simulation Experiment

To explore the real-world impact of doppelgänger effects, I adopted the K-Nearest Neighbor (KNN) method to model a pulmonary tuberculosis dataset (Dou et al., 2021) with data doppelgänger and without data doppelgänger respectively, which led to the results as shown in Figure 1. According to Table I, data doppelgänger has a considerable impact on Machine Learning. It can be found out that the performance of the same model varies significantly in the presence and absence of data doppelgänger, which implies the possibility that doppelgänger effects can make the model perform falsely well.

Table I. The Model Performance

	Dataset without the data doppelgänger	Dataset with the data doppelgänger
Accuracy	76.67%	93.33%
Precision	70.26%	95%
Recall	76.62%	93%
F1 score	0.72	0.93

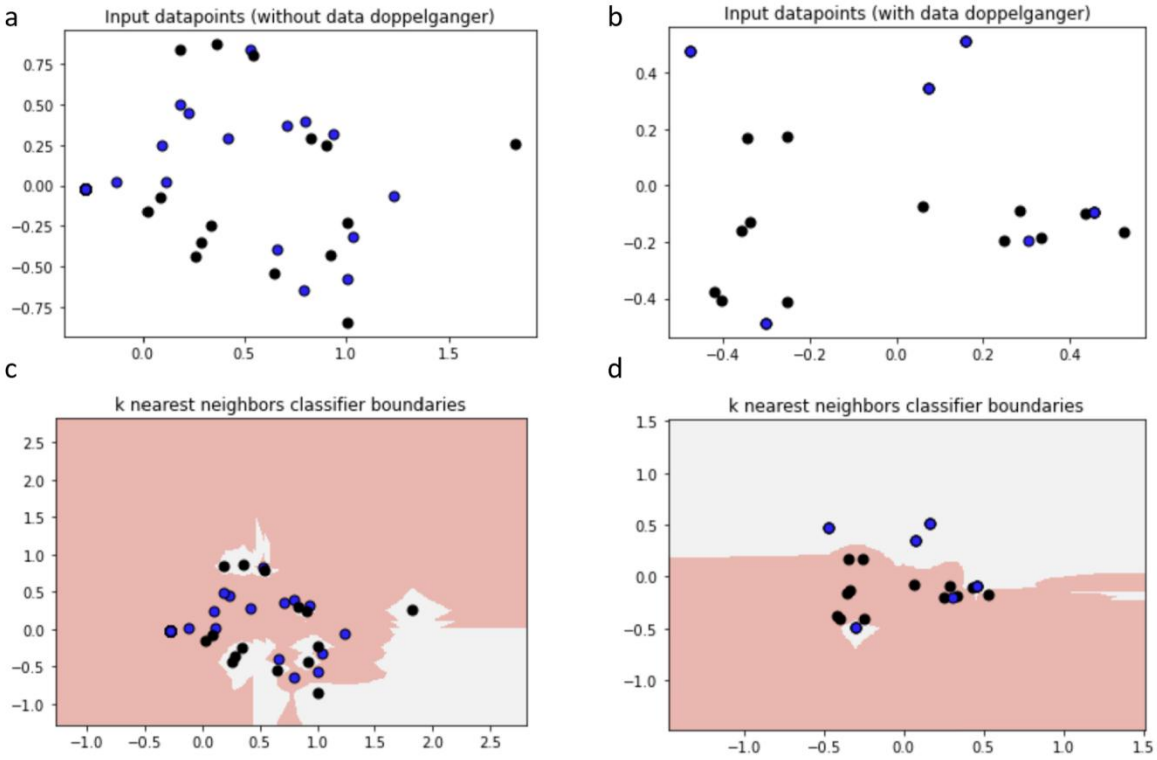

Figure 1. The simulation experiments conducted to explore the impact of the doppelgänger effects; (a) and (b) The data distribution after dimension reduction through principal components analysis; (a) The dataset with data doppelgänger; (b) The dataset without data doppelgänger; (c) and (d) The specific classifier boundaries after the classification; (c) The dataset with the data doppelgänger; (d) The dataset without data doppelgänger.

Discussion

In general, there are more cases of doppelgänger effects in biomedical data due to the positive correlation exhibited by transcriptional profile of genes (Venet, Dumont, & Detours, 2011). However, I consider that doppelgänger effects are not unique to

biomedical data. It is common that highly similar data are generated in social life. For example, we will also encounter the doppelgänger effect when using social media data to predict the box office. We can analyze the emotional tendency of users to unreleased films through natural language process technology before the transformation of it into the viewing needs of users. In data collection, however, there are plenty of similar reviews left. In-depth microblog retweets and active comments are the main contributors to data doppelgänger. If these data doppelgängers are left unprocessed during the analysis, the erroneous prediction of overly high or too low viewing requirements may result, thus causing problems when there are new datasets for the model. Therefore, doppelgänger effects are not unique to biomedical data.

To eliminate the doppelgänger effect from Machine Learning models in the context of health and medical science, I propose to adopt the "machine + human" approach, which means it is indispensable to improve machine learning models and manually process data. For example, different Machine Learning (ML) models were applied in this paper to confirm data doppelgängers act as functional doppelgängers, with inflationary effects produced. It was also indicated that not all models are equally affected. The K-nearest neighbor (KNN) model shows a more apparent linear relationship between performance expansion and split-dose than other models. In this part, the variation in sensitivity of the model to the data doppelgängers illustrated, meaning that the effect may be weakened to some degree by changing the model. Take the KNN algorithm as an example. The mechanism of this algorithm, to some extent, determines the higher sensitivity of this model to data doppelgängers. The core idea of the KNN algorithm is to identify a sample and its other k most similar samples in the dataset, with the category of that sample judged by the most common category in these k samples. That is to say, the model results can be significantly affected by the existence of multiple highly identical doppelgängers pairs. Therefore, it is easy to explain why the KNN model exhibits a more evident linear relationship between performance inflation and doppelgänger dosage. Thus, it is possible for the doppelgängers effect to diminish because of the switch to some algorithms with a low sensitivity to data doppelgängers. However, the training of model is inseparable from

the data, which makes it necessary to process the data doppelgängers. In my view, performing data stratification may be an effective solution to carrying out intervention because new associations could be revealed by data stratification (Wei et al., 2021). Before data collection, it is essential to consider what information about the data sources might affect the results. On this basis, the performance can be compared between the different strata and the strata with poor model performance pinpoint gaps in the classifier. This gap may be the key to the problem.

Reference:

Dou, W., Liu, Y., Liu, Z., Yerezhepov, D., Kozhamkulov, U., Akilzhanova, A., Dib, O., Chan, C.-K. (2021). An AutoML Approach For Predicting Risk Of Progression To Active Tuberculosis Based On Its Association With Host Genetic Variations. *In 2021 10th International Conference on Bioinformatics and Biomedical Science (ICBBS 2021), October 29-31, 2021, Xiamen, China.* ACM, New York, NY, USA, 10 Pages. https://doi.org/10.1145/3498731.3498743

Venet, D., Dumont, J. E., & Detours, V. (2011). Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. *PLOS Computational Biology, 7*(10), e1002240. doi:10.1371/journal.pcbi.1002240

Wang, L. R., Wong, L., & Goh, W. W. B. (2021). How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*. doi:https://doi.org/10.1016/j.drudis.2021.10.017

[1] Wei, J., Cheng, L., Han, P., Zhu, Y., & Huang, W. (2021). Decision Tree-Based Data Stratification Method for the Minimization of the Masking Effect in Adverse Drug Reaction Signal Detection. *Applied Sciences, 11*(23), 11380.