Analyzation of doppelgänger effects in health and medical sciences

Machine learning is a multi-field cross-technology. Since machine learning models can improve the efficiency of medicine discovery, it has improved in the discovery of, for example, anticancer drug candidate EXS21546 and several ML-identified drugs and drug combinations for the treatment of COVID-19 in recent years [1]. Therefore, machine learning is widely used in the biomedical field. However, the application of this technology in the biomedical field still needs improvement, as it is difficult to remove the data doppelgängers [2].

This report will first define data doppelgängers and doppelgänger effects and analyze them from a quantitative perspective. After that, the report will explain why doppelgänger effects do not exist only in biomedical. Finally, the report will illustrate how doppelgänger effects can be checked and avoided in the practice and development of machine learning models in the health and medical sciences.

Data doppelgängers refer to the training and validation set of machine learning that is highly similar to each other, while they should be independently derived [3]. Data doppelgängers cause models to perform well regardless of how they are trained.

Doppelgänger effects will be observed when a classifier falsely performs well due to data doppelgängers [3]. In other words, the system's performance is exaggerated because the system is evaluated on a test set that is highly similar to the training set. This will lead to the that even if the selected feature is random, there will be some test data with a good performance given the particular training data [2]. From a quantitative perspective, doppelganger effects may occur due to the data. Assuming that a number of data from multiple sources is integrated, then it is likely to have the same individual represented by different records in each source, which leads to a doppelganger effect. Besides, If the data is suddenly missing or incomplete, the system will then be difficult to identify and match records. And lead to the same result. For some more complex systems, for example, to protect each person's identity, the system can use different identifiers for different datasets, leading to a doppelganger effect.

Doppelganger effects are not unique to biomedical data. From the perspective of data science, doppelganger effects mean that two or more records (training set) in the data set point to the same individual (validation set), but have different identification information. This can occur in any dataset that includes identifying information, such as financial, criminal, or credit records. In biomedical data, doppelganger effects can occur in evaluating existing chromatin interaction prediction systems and predicting protein function [4]. Although there are many data doppelgängers and doppelgängers effects in biomedical data, they have not been characterized until now [4].

The accuracy of the model on the validation data is often used to evaluate the performance of the machine learning model. This method is only effective when validation data is different from the training data, which is usually assumed to be true when analyzing. However, this assumption may not hold when the doppelganger effect

occurs [2].

Wang, Wong, and Goh attempt several methods to avoid the doppelganger effect in the practice and development of machine learning models for health and medical science [2]. Their methods are classified into three categories: one is not feasible, the other is limited, and the third has not been attempted.

Firstly, the non-feasible methods category, which includes using ordination methods or embedding methods to show how instances are distributed in reduced-dimensional space [2]. However, it is unfeasible because data doppelgängers may be undistinguishable in reduced-dimensional space. Moreover, the dupChecker method was used in the early research. Duplicate samples are identified by comparing the MD5 fingerprints. However, this method is not widely used because it does not detect the true data doppelganger [5]. Additionally, Cao and Fullwood raise the idea of splitting training and testing data based on individual chromosomes. However, this is hard to achieve because it relies on prior knowledge and good quality contextual/benchmarking data [6].

Secondly, the category of methods with limitations. The pairwise Pearson's correlation coefficient (PPAC) generates the PPAC value of different data pairs. The sample pairs with high PPCC values will be PPCC data doppelgängers [2]. However, PPAC cannot determine which one between the pair is the original data. Meanwhile, recent research shows that the results do not constitute true data doppelgängers. Then, Lakiotaki, Vorniotakis, Tsagris, Georgakopoulos and Tsamardinos used doppelgangR to identify doppelgängers. And removed PPCC data doppelgängers to reduce their effects, successfully improving data doppelgängers [7]. However, there exist some limitations to this method. Take the KNN model as an example of removing highly correlated variables.

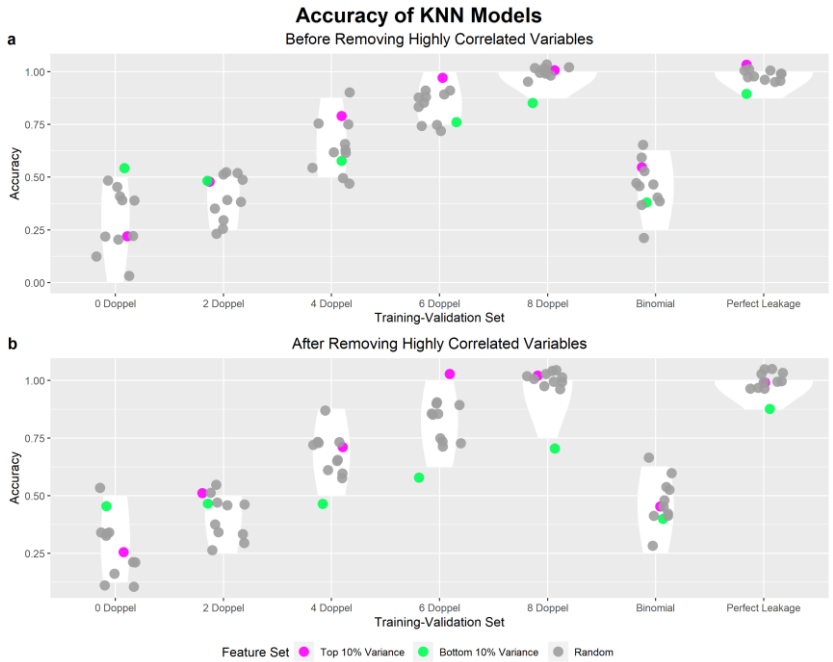

Figure 1: Accuracies of KNN models before (a) and after (b) removal of highly

correlated variables [2]. X-axis: Type of validation set: “i Doppel” refers to a validation set with i number of PPCC data doppelgängers in the training set (where i = 0, 2, 4, 6, 8), “Binomial” refers to the accuracies generated by 12 (negative control), “Perfect Leakage” refers to a validation set with 8 duplicates with the training set (positive control). Y-axis: Accuracy of machine learning models on a validation set of 8 samples. Legend: “Top 10% Variance” refers to the feature set comprising of proteins of the highest variance at 10% of the total number of proteins in the dataset, “Bottom 10% Variance” refers to the lowest variance at 10%, “Random” refers to randomly select proteins at 10% [2].

The inflationary effects of PPCC data doppelgängers remain after removing these variables. Hence, the removal of highly correlated variables is not a viable method of PPCC data doppelgänger amelioration [2]. Meanwhile, this method cannot support small data sets with a high proportion of PPCC data doppelgängers [7]. Finally, one more suboptimal solution is placing PPCC data doppelgängers in either the training or testing set. There will be no doppelgänger effect if one of the datasets is empty, and the accuracy will become 0.5 [2]. By the way, this suboptimal method is not recommended.

Thirdly, the category of the methods that have not been attempted. The first method is to use meta-data, which may anticipate PPCC value ranges for the condition in which doppelgängers cannot exist [2]. Even though the samples rise from the same class but different individuals, it will be able to recognize recessive doppelgängers and separate them into different sets with the support of meta-data [2]. The second method is data stratification. Data will be stratified into strata according to similarities and evaluate model performance on each stratum separately [2]. More importantly, strata with poor model performance pinpoint gaps in the classifier. The third method is to perform highly robust independent validation checks, which can inform the objectivity of the classifier [8].

The above are feasible and pending methods proposed by Wang, Wong, and Goh. From a data science perspective, there are some possible ways to help check and avoid data doppelgängers. The first method is Data Cleansing, which includes tasks such as validating data and removing duplicates. The second method is Record Linkage, which identifies and links records that refer to the same individual across different sources. The last method is to use Blockchain. Blockchain technology can create a decentralized and tamper-proof database, which can help ensure the integrity and authenticity of the data. This can be useful for resolving doppelganger effects. All the methods can also be used in combination to effectively avoid and check for doppelganger effects, and ensure the accuracy and integrity of the data.

In conclusion, data doppelgängers and doppelgänger effects are the challenges that need to be improved in machine learning of health and medical science. Besides, Doppelganger effects are not unique to biomedical data. Therefore, it can be checked and avoided from two perspectives, biomedical science and data science. Meanwhile, although the method raised by Wang, Wong, and Goh may have application limitations, the basic design and the recommendations methods are meaningful.

Reference

[1] Y. Zhou, F. Wang, J. Tang, R. Nussinov, and F. Cheng, "*Artificial Intelligence in covid-19 drug repurposing*," The Lancet Digital Health, vol. 2, no. 12, 2020.

[2] L. R. Wang, L. Wong, and W. W. Goh, "*How doppelgänger effects in biomedical data confound machine learning,*" Drug Discovery Today, vol. 27, no. 3, pp. 678–685, 2022.

[3] S. Y. Ho, K. Phua, L. Wong, and W. W. Bin Goh, "*Extensions of the external validation for checking learned model interpretability and generalizability,*" Patterns, vol. 1, no. 8, p. 100129, 2020.

[4] W. W. Goh and L. Wong, "*Turning straw into gold: Building robustness into gene signature inference,*" Drug Discovery Today, vol. 24, no. 1, pp. 31–36, 2019.

[5] Q. Sheng, Y. Shyr, and X. Chen, "*Dupchecker: A bioconductor package for checking high-throughput genomic data redundancy in meta-analysis*," BMC Bioinformatics, vol. 15, no. 1, 2014.

[6] F. Cao and M. J. Fullwood, "*Inflated performance measures in enhancer–promoter interaction-prediction methods*," Nature Genetics, vol. 51, no. 8, pp. 1196–1198, 2019.

[7] K. Lakiotaki, N. Vorniotakis, M. Tsagris, G. Georgakopoulos, and I. Tsamardinos, "*BioDataome: A collection of uniformly preprocessed and automatically annotated datasets for data-driven biology,*" Database, vol. 2018, 2018.

[8] S. Y. Ho, K. Phua, L. Wong, and W. W. Bin Goh, "*Extensions of the external validation for checking learned model interpretability and generalizability*," Patterns, vol. 1, no. 8, p. 100129, 2020.