1. Introduction

1.1 Definition

Data doppelgängers mean that samples are very similar to each other. While this phenomenon causes machine learning to perform well falsely, we may state that this data doppelgängers generate doppelgängers effect [1]. Therefore, not every data doppelgänger will cause such an effect, making the concrete definition of its concept difficult. In clinical genomics, several doppelgängers' effects have been identified based on the expression data and clinical annotations [2-3]. Feature reduction methods such as PCA and t-SNE may help us understand whether the doppelgänger exists based on the reduced dimensional space. However, there is no evidence to show that data doppelgänger will be distinguished in the lower dimension.

1.2 Identification

Although other methods such as dupChecker and PPCC may discover the data doppelgänger, either of them can indicate the link between functional doppelgänger and data doppelgänger or detecting the true data doppelgängers. In this article, the author applied the PPCC method to detect the possible functional doppelgänger from constructed benchmark scenarios, trying to figure out the link between data doppelgängers and effect. The result indicates that the more PPCC doppelgängers data exists, the more inflation is caused in the ML model. While different kinds of ML models may not be equally affected, for instance, KNN and naïve Bayes are more sensitive to the doppelgängers effect than logistic and decision trees.

1.3 Amelioration

Previously, people tried to put all the doppelgängers data into the training dataset, avoiding abnormal performance in the testing dataset. However, this method cannot be the most optimized due to the limitation of model generalization and data size. Also, direct removing the doppelgängers sample may cause a limited size of the training dataset, causing bad performance.

Although removing the doppelgängers effect can be elusive, the author of this article provides some possible methods for amelioration.

(1) Using meta-data as a guide to constructing the training and validation dataset.

(2) Applying the validation on data stratification. Testing the performance of model on each stratum such as non-PPCC data doppelgängers and PPCC data doppelgängers.

(3) Using robust validation methods such as divergent validation.

2. Answers to the questions

(1) Whether you think doppelganger effects are unique to biomedical data?

I don't think the doppelganger effects are unique in biomedical data. Although many studies are trying to figure out the doppelganger within genomic expression datasets, it does exist some instances confounding by the doppelganger effect. Taking pattern recognition as an example, due to the limitation of figures, some samples in the training dataset may share similarities with the data in validation, causing unexpected performance. For instance, people want to classify the 'Wolves and Dogs' based on several figures. Although the ML model performs well in both training and validation datasets with no evidence for overfitting, it performs worse while taking into the real condition. To figure out the problem, people find that most of the figures of wolves had the background of icy land or desert, while domestic dogs are stayed in the city or home, far from the natural environment. This phenomenon can be explained by the data leakage or doppelganger due to the similarity in samples.

(2) How you think it can be avoided in the practice and development of machine learning models for health and medical science?

In my opinion, I think three ways can be applied to reduce the effect of doppelganger.

1. **Testing the ML model in other datasets based on prior knowledge in certain area**. Although the author mentions that prior knowledge can be difficult to pursue, sometimes it can be the most direct way to remove the effect. Taking lncRNA (Long noncoding RNA) classification as an example, scientists are trying to design one ML model to classify lncRNA and coding RNA. However, lncRNA shares a similar structure in the species' transcriptome. They may share the same putative ORF (Open reading frame) or even the same fragments. Therefore, in one species, it may have several lncRNA samples that share a similar structure. To prevent the doppelganger

effects, scientists will train the ML model mainly based on the RNA databases in humans. Then, applying this model to other databases of different kinds of species such as a rat, insects, or fish. Taking the article on lncRNA-MFDL as an example, the author trained the lncRNA-MFDL model on the human database and compared the performance with the previous methods – CPC and CNCI [4]. If doppelganger effects in human RNA databases do influence the performance of the model, it will have poor performance on other databases.

2. **Using ensemble learning methods such as bagging.** The bagging method will divide the training dataset into several small sub-datasets. Then, feeding each sub dataset to each weak learning model. Each model will train independently based on its own datasets. If the data doppelganger is not popular in the dataset, the random division of bagging may prevent each model to receive a similar sample. Also, the ensemble learning will take the average value as the final output which means that it is robust enough when some weak models are still influenced by the data doppelganger.

3. **Adversarial Machine Learning in pattern recognition**. Sometimes the samples are rare and hard to be recollected. To prevent data doppelgangers, we may invent some novel samples. For instance, adding some noise to the original figure or inverting the color of the target sample. Then, figuring out whether the model can still classify these samples or performances much worse than before. This method can be applied in the 'Wolves and Dogs' classification task. Due to the similarity between the background of each wolf's sample, adversarial machine learning can change the color of the background or even remove the background. Then, testing whether the model will perform well or worse.

Reference

[1] S.Y. Ho, K. Phua, L. Wong, W.W.B. Goh, Extensions of the external validation for checking learned model interpretability and generalizability, Patterns 1 (2020) 100129

[2] Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Van de Vijver MJ, Bergh J, Piccart M, Delorenzi M. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst. (2006) Feb 15;98(4):262-72.

[3] Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A. (2005) Sep 20;102(38):13550-5.

[4] Fan, Xiao-Nan, and Shao-Wu Zhang. "lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning." Molecular BioSystems 11.3 (2015): 892-897.