Report for *How doppelgängers effects in biomedical data confound machine learning*

Machine learning models are wildly used in the drug industry to efficiently discover drug development and identify existing drugs while dramatically decreasing the cost of the process. However, the existence of data doppelgängers decreases the performance of ML models on training. Data doppelgängers are independently derived data that are very similar to each other[1]. The effect of doppelgängers may leads to a well-performed training result. Therefore, the identification of data doppelgängers is necessary.

Data doppelgängers exist in many biological cases. For example, some proteins and their ancestor proteins have similar sequences, which may be presumed to be similar in function[1]. Moreover, some doppelgängers as duplicate pairs are discovered in database. “There are duplicates originated between the TCGA datasets and datasets of institutions that contributed samples to the TCGA project”[2]. In addition, the similarity between training and testing sets may cause doppelgängers effect as well. Therefore, the identification of data doppelgängers has a significant meaning.

In this paper, it provides some methods used to identify certain data doppelgängers. DupCheckers identifies duplicate samples by comparing the fingerprints of their CEL files[3]. It detects essentially replicates but not similar independently derived doppelgängers. The other method is Pairwise Pearson’s Correlation Coefficient(PPCC) which captures relations between sample pairs of

different datasets[1]. A high value of PPCC implies that there are data doppelgängers between pairs of samples. Data doppelgängers are detected when valid sample pairs with PPCC values are greater than the maximum PPCC of negative sample pairs.

Notice that the machine learning model is learned to form training sets. It is reasonable to check if PPCC data doppelgängers of training and validation effect on the performance of machine learning. The result from comparing the distribution of PPCC values from patients with a class labeled shows that PPCC data doppelgängers confound machine learning outcomes.

Since the data doppelgängers produce inflationary effects on machine learning, the management of doppelgängers effects is necessary. Besides removing data doppelgängers directly, in this paper, it is recommended to use the meta-data to cross-check. Also, stratifying data into strata of different similarities and performing models on each stratum separately[1] compared to the performance of different strata from the real world.

In general, the accuracy of results from the machine learning model relies on the independence of training sets. Data doppelgängers are commonly exist in our data. It is unfortunate that the doppelgängers effects is hard to avoid. The most basic way I think to avoid doppelgängers effect is to cross-check the similar resource of data before implementing(i.e check the origin of the data). Indeed, data doppelgängers have an inflation effect on training machine learning models in the area of biomedical. I believe that the doppelgängers effects are not unique to biomedical but also in other areas.

It is known that government, organizations, and companies create an impression of a crowd that supports a specific opinion by producing fake news, comments, reports, and articles[4]. In the article *Hi doppelgängers : Towards Detecting Manipulation in News Comments. Retrieved from*, it demonstrates how to detect doppelgängers in news comments. Although it is not a method used to reduce the doppelgängers effect on a model, it may be used to distinguish doppelgängers data in linguistics content. For example, to avoid doppelgängers effect when collecting data about fake news, this method could be used to remove similar content or resource of fake news coming from one organization when detecting doppelgängers. It may help to reduce doppelgängers effect since amount of fake news generated from one organization might have lots of similarities.

Reference

1. Wang. L., Wong. L & Goh . W. *How doppelgängers effects in biomedical data confound machine learning.* Retried from *https://doiorg/10.1016/j.drudis.2021.10.017*

2. Waldron. L et al.(July 2016) Journal of the National Cancer Institute. *The doppelgängers effect: hidden duplicates in databases of transcriptiome profiles*. Retrived from ncbi.nlm.nih.gov/pmc/articles/PMC5241903/

3. Q. Sheng, Y. Shyr, X. CHen, DupChecker. A bioconductor package for checking high-throughput genomic data redundacy in meta-analysis, BMC Bioinform 15 (2014) 323.

4. Pennekamp. J, Henze. M & Hohlfeld .O. Hi doppelgängers : Towards Detecting Manipulation in News Comments. Retrieved from Martinhenze.de/wp-content/papercite-data/pdf/phhp19.pdf