The doppelganger effects in biomedical data confound machine learning

1. Introduction

When a classifier falsely performs well because of the presence of highly similarity between training and validation sets, we say that there is an observed doppelganger effect.[1]

The performance of ML models is usually evaluated by the accuracy of the model on the validation set. The validation set used for evaluation must be independent from the training set, which is a broad consensus in the field of machine learning. However, due to the prevalence of the doppelganger effects in the biomedical data set, this broad consensus may not hold true. The inflationary effect caused by the doppelgangers might lower the classification accuracy of machine learning model, which will directly affect the efficiency of machine learning in the field of drug development, so it is vital to check the potential doppelgangers in the data before training and validation data split.

2. The prevalence of doppelganger effects

The recent studies have proved the abundance of doppelgangers in biological data. However, in my opinion, the doppelganger effects are not unique to biomedical data.

1) The doppelganger effects in biomedical data

Doppelganger effects have been observed in modern biomedical research. Cao and Fullwood found that the test sets for a existing chromatin interaction prediction system shared a high degree of similarity to training sets.[2] Same situation happened in many feilds in bioinformatics: in protein function prediction[3] and drug discovery[4]. Even though the biomedical data science community have awared of the doppelganger problems, there isn't a standard solution for eliminating the similarity between test and training data to avoid doppelganger effects.

2) Doppelgangers are not unique for biomedical data

As long as the training set and testing set of the ML model are derived from the same data set, theoretically, there could be a certain degree of similarity between the training and testing set which means the existence of potential doppelgangers.

For example, in the field of handwritten digit recognition, many researchers' machine learning models achieved an error rate of less than 1% on the mnist data set. However, some images in the training and test sets of mnist have extremely high similarities, which can lead to falsely high accuracy. When these ML models are tested on other handwritten digital data sets, they often fail to reach the prediction accuracy of the evaluation on mnist test set. I take it as a specific manifestation of doppelganger effects, the doppelganger effect therefore is not unique for biomedical data.

3. The implications of data doppelgangers on ML

The pairwise Pearson's correlation coefficient(PPCC) is taken as a measurement of relations between sample pairs of different data sets.[5] The samples pairs are divided to 3

three types based on the similarity of their patient and class: positive pairs(same patient same class), valid pairs(same class different patient) and negative pairs(different classes). The sample valid pairs are usually identified as possible doppelgangers.

High PPCC value indicates the doppelgangers. The study shows that presence of PPCC data doppelgangers in both training and validation data results in good ML performance, the more PPCC doppelganger pairs presented in sets, the better the ML performance.

The doppelgangers which could cause doppelgangers effects are termed functional doppelgangers. The study confirms that if PPCC data doppelganger act as functional doppelganger, the inflationary effect caused by it is close to data leakage.

4. The method to overcome the doppelganger effect

Cao and Fullwood called for a good standard in ameliorating data doppelgangers. [2] However, it's very hard to do practically because it predicates on the existence of prior knowledge and good quality contextual data.

1) The recommendation proposed by this study

The first recommendation is to perform careful cross-checks using meta-data as a guide. The study used the meta-data in RCC[6] for constructing cases which allow us to classify the the sample pairs. As Fig.1 shows below, doppelganger can not exist in negative pairs(different class), leakage exists in positive pairs(same-patient and same class) and valid pairs(same class different patient) may consists of doppelgangers. With this information from meta data, the potential doppelgangers can be identified.

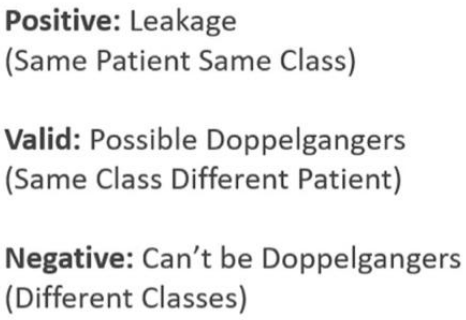

Figure 1 Naming convention for different type of sample pairs

The second recommendation is to perform data stratification. Instead of evaluating model performance on whole data, the study suggested to stratify the data into different layers based on the similarity and evaluate the model performance on each layer separately.

The third recommendation is to perform extremely robust independent validation checks involving as many data sets as possible.[1] Although this method is not a direct hedge against data doppelganger, divergent validation techniques could no doubt improve the generalization ability of the model.

2) My idea on how to mitigate doppelganger effect

As far as I am concerned, the way to avoid the doppelganger effect consists of two steps. The first step is functional data doppelgangers identification, the second step is taking measures to process the detected doppelganger to reduce the effect of doppelganger.

Few doppelganger identification methods have been proposed by the formal studies. For example, doppelgangR was used for the identification of doppelgangers, PPCC data doppelgangers could be removed to mitigate their effects.[7] But this approach does not work on small data sets with a high proportion of PPCC data doppelgangers. So, it's crucial to exploring a methods of functional doppelganger identification that do not rely heavily on meta data.

Two methods can be taken to process the detected functional doppelgangers. The first method is remove those have been identified. This method is not suitable for the dataset with high percentage of doppelganger. The second method is assort all potential doppelgangers into either training or validation sets.

Besides, training the model with different datasets which contains varied features and performing independently validation as much as possible will theoretically mitigate the effect of doppelganger.

5. The doppelganger effect in datasets of cancer transcriptome profile

This is an interesting example of doppelganger effect in the RNA sequencing data I found in Waldron's study.[8] In today's cancer genomic anlysis, the investigators frequently share or re-use specimens in later studies. However, these wide reused public datasets proved to contain lots of duplicate expression profiles(duplicate RNA sequences) which falsely inflated prediction accuracy and confidence in differential expression.

1) The cause of doppelganger effect in RNA sequencing data

Publicity available human genomic data is normally summarized at a level that cannot be identified uniquely to protect patient privacy. Cancer transcriptomes undergo alterations that are much more difficult to identify in summarized form. So, re-use of tissue specimens is widespread in clinical genomic studies which finally cause doppelganger effects.

2) The method to identify doppelganger effects

In my opinion, it takes three steps to identify the duplicate RNA sequences between two dataset. The first step is to choose transcript identifiers available in both sets as features of each sample. The second step calculating Pearson's Correlation Coefficient(PCC) between every sample in one dataset against every sample in the other dataset. The last step is identifying the duplicate RNA samples(doppelgangers) based on the value of PCC between two samples.

6. 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] F. Cao, M.J. Fullwood, Inflated performance measures in enhancer–promoter interaction-prediction methods, Nat Genet 51 (2019) 1196–1198.

[3] M.N. Wass, M.J. Sternberg, ConFunc: functional annotation in the twilight zone, Bioinformatics 24 (2008) 798–806.

[4] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, R.K. Tekade, Artificial intelligence in drug discovery and development, Drug Discov Today 26 (2021) 80–93.

[5] Waldron L , Riester M , Ramos M , et al. The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles[J]. Journal of the National Cancer Institute, 2016.

[6] Guo, Tiannan, et al. "Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps." Nature medicine 21.4 (2015): 407-413.

[7] Kleanthi L , Nikolaos V , Michail T , et al. BioDataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology[J]. Database The Journal of Biological Databases and Curation, 2018.

[8] Waldron L , Riester M , Ramos M , et al. The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles[J]. Journal of the National Cancer Institute, 2016.