Report for Q3

Introduction

Overlaps between train set and test set would easily lead to mistakes of ML models' performance evaluation. Hence it is important to divide dataset carefully. However, similarities between train set and test set not only come from division error. Indeed, doppelganger effects, which is used to describe that similar but independent derived data gives rise to fake good testing result of a ML model regardless of its training performance, present great prevalence in biomedical field. Considering that Machine Learning significantly prompts the biomedical development like increasing the efficiency of drug discovery, the existence of doppelganger would confound biomedical researches based on ML. Therefore, it is important to understand and control doppelganger effects. This report focuses on the paper *How doppelgänger effects in biomedical data confound machine learning* talking about the problem of doppelganger effects in bioinformatics.

Summary

This part is about the summary of abundance of doppelganger effects and the influence of it, as well as how to mitigate it.

1. Abundance & Identification

   The characteristics of the observation objects of bioinformatics make doppelganger effects easily to generate. For example, some proteins observed for protein function predictions presents similarity because they inherit from the same ancestor protein. What's more, during drug discovery, there are some similar modules owning different activities, which is important but hard to differentiate.

   To identify doppelganger effects, it essential to notice that data doppelganger are not necessarily distinguished in reduced-dimensional space and the data should be independent with each other and similar be chance. Thus, it's suitable to use PPCC methods, which identify data doppelganger from the prospective of capturing relations between sample pairs of different data sets.

2. Confounding Influence

   The confouding effects of data doppelganger are kind of similar to data leakage. The test results are usually inflationary but fake. The overlaps in data is also a kind of information loss, which means the information that could be learned by a ML model is reduced. Just like students do some homework exercises repeatedly. The score of their homework might look great, but it is not a incredible evaluation of their study performance.

   From a **quantitative view**, the degree of doppelganger effects are regular. From the accuracy tendency graphs in the paper, we could find that the accuracies increase while the number of data doppelganger in the training-validation set go from 0 Doppel to 8 Doppel. But when all data doppelganger are input into the training-validation set, the accuracies will decrease.

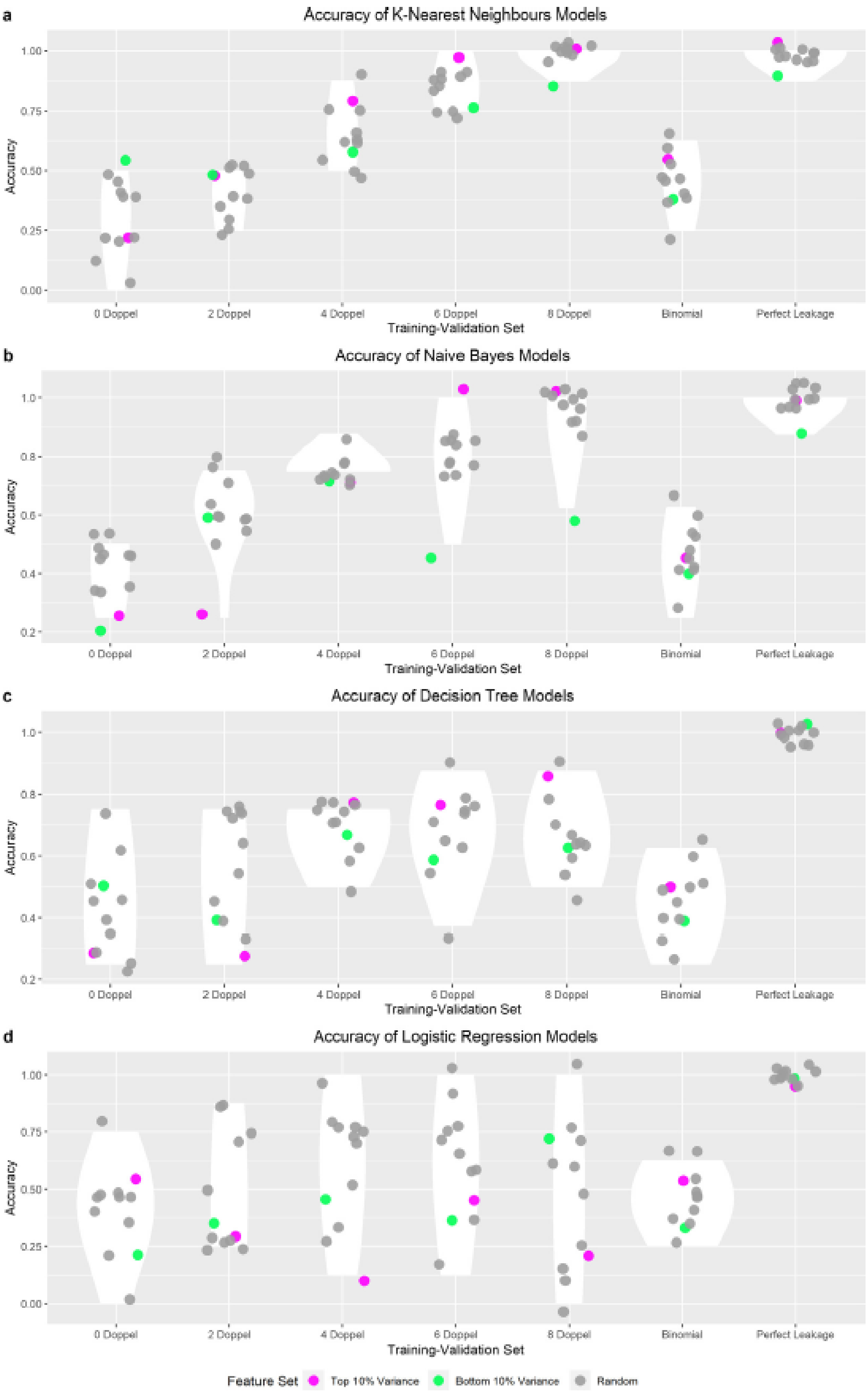


Figure 1: Accuracy of different methods under various data doppelganger in training-validation set[1]

3. Mitigation

There are several methods mentioned in the paper to eliminate doppelganger effects.

- Place all doppelganger in training set

  Disadvantage: Models will lack knowledge in case of the size of the training set is fixed.

- Split train set and test set based on individual chromosomes

    Disadvantage: Requiring prior knowledge and goof quality contextual/benchmarking data, which is hard to practice.

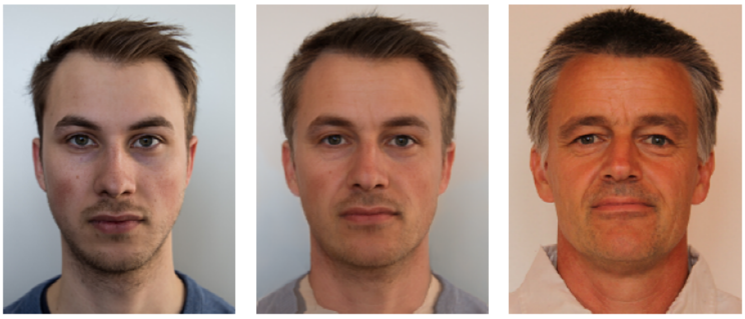
  - Remove data doppelganger directly

    Disadvantage: Leading to significant reduction in sample size, but abundant training data is important to the performance of ML model training.

  All the possible mitigation methods have its own shortcomings. So the paper recommends to use cross checks or validation checks to ensure the robustness of the model. Or changing test strategies. Data stratification could also archive the testing effects of the whole test dataset.

Response

In fact, **I don't think doppelganger effects are unique in biomedical data**, although they are prevalent in this field. It is actually a probability problem. If the probability of generating similar, or even the same data is big enough, then doppelganger effects will appear. For instance, doppelganger effects could also appear in the research on human face images. We often regard facial features as a unique mark of a human. However, it is not difficult to find two persons look similar to each other. Alexander Rottcher et al utilize data doppelganger in **human face images** to realize face morphing attack, which is a threaten to the security of face detection nowadays.



(a) Subject A. (b) Morph. (c) Subject B.

Figure 2: An example of morphed images[2]

Since doppelganger effects are so widespread and the appearance of doppelganger data is hard to avoid, it is much necessary for us to understand it and learn to control it. According to the paper, **I think the strategies of checking and data stratification are more reasonable for managing doppelganger effects**, because they have less side effects.

Conclusion

Doppelganger effects are prevalent in bioinformatics, and both the similarity and proportion of data doppelganger could influence this effect. During the training and testing periods of ML models, the performance of the model evaluated by validation data could be misled by doppelganger effects. Therefore, it is necessary to understand the doppelganger effect and manage it. The best managing strategy now is to check potential doppelganger before running on the data.

Reference

[1] Li Rong Wang, Limsoon Wong, Wilson Wen Bin Goh, How doppelgänger effects in biomedical data confound machine learning, Drug Discovery Today, 2021.

[2] A. Röttcher, U. Scherhag and C. Busch, "Finding the Suitable Doppelgänger for a Face Morphing Attack," 2020 IEEE International Joint Conference on Biometrics (IJCB), 2020, pp. 1-7, doi: 10.1109/IJCB48548.2020.9304878.