Report of How doppelgänger effects in biomedical data confound machine learning

1. Relevant background

Facts about data doppelgängers

Classification models based on ML and AI has been increasingly used in drug discovery to speed up drug development, their accuracy depends on how properly they are trained and tested. But models trained and validated on data doppelgängers might falsely perform well regardless of the quality of training, which we can say there is an observed doppelgänger effect.

Data doppelgängers was earlier stated to be when samples appear similar across their measurements, but this may not guarantee a doppelgängers effect. Thus data doppelgängers that also generate a doppelgängers effect which can confound ML outcomes are termed functional doppelgängers. Data doppelgängers and its accompanying effects are poorly documented and understood and uncommon to be checked. Though there are several proposed methods of identifying data doppelgängers, they are yet not generalizable or robust enough. Hence, it is imperative to investigate the nature of data doppelgängers and propose improved methods for doppelgänger identification.

There are abundant of data doppelgängers in biological data, for instance, in established fields of bioinformatics and in drug discovery. Though the biomedical data science community appears to be increasingly aware of such data doppelgängers, procedures for eliminating or minimizing similarity between test

and training data still do not constitute standard practice before classifier evaluation.

Methods to identify data doppelgängers

1) The ordination and embedding methods coupled with scatterplots to see the distribution of samples in reduced-dimensional space. The study find such method to be unfeasible for data doppelgängers are not necessarily distinguishable in reduced-dimensional space.
2) The dupChecker identifies duplicate samples by comparing the MD5 fingerprints. Thus it does not detect true data doppelgängers that are independently derived samples that are similar by chance.
3) The pairwise Pearson's correlation coefficient (PPCC) captures relations between sample pairs of different data sets. Although reasonable and intuitive, it never conclusively made a link between PPCC data doppelgängers and their ability to confound ML tasks.

2. Writing purposes

A. To better understand the level of similarity between suspected functional doppelgängers and the acceptable proportion of functional doppelgängers in the validation set.

B. To find ways to mitigate the doppelgänger effect.

3. Findings

This study uses the renal cell carcinoma (RCC) proteomics data to construct benchmark scenarios and uses the PPCC method for identifying and finds out that

PPCC has meaningful discrimination value.

After identifying PPCC data doppelgängers in RCC, the result points toward a dosage-based relationship between the number of PPCC data doppelgängers and the magnitude of the doppelgänger effect, and confirms that PPCC data doppelgängers act as functional doppelgängers, producing inflationary effects similar to data leakage.

4. Solutions

Possible solutions

1) Place all doppelgängers in training set could eliminate the doppelgänger effect, but it is not suboptimal solution.
2) Use the PPCC outlier detection package doppelgangR to remove PPCC data doppelgängers to mitigate their effects, but it does not work on small data sets with a high proportion of PPCC data doppelgängers such as RCC.
3) Remove variables contributing strongly toward data doppelgängers effects, but observe no change in the inflationary effects of the PPCC data doppelgängers.

Recommended solutions

1) Perform careful cross-checks using meta-data as a guide: use the meta-data for constructing negative and positive cases, then anticipate PPCC score ranges for scenarios in which doppelgängers cannot exist, thus we would be able to identify potential doppelgängers and assort them into

either training or validation sets to prevent doppelgänger effects.

2) Perform data stratification: stratify data into strata of different similarities instead of evaluating model performance on whole test data.

3) Perform extremely robust independent validation checks involving as many datasets as possible.

Future work

Explore other methods of functional doppelgänger identification that do not rely heavily on meta-data in order to identify functional doppelgängers directly. Further pairing this approach with PPCC subsequently may allow researchers to discern the doppelgänger partners of test set samples in the training set.