Answer to Question 3

How doppelgänger effects in biomedical data confound machine learning

Wang, L. R., Wong, L., & Goh, W. W. B. (2022), Drug discovery today, 27(3), 678–685.

1. Introduction

Due to the growing enormity and complexity of biomedical data, machine learning (ML) is becoming an increasingly important analytical tool. Through sophisticated algorithms, ML can analyze large-scale and heterogeneous datasets to discover meaningful patterns, which otherwise would be very difficult and inefficient to identify manually. Enormous amounts of data are used as inputs to these simulation models and hence the success of the models is heavily dependent on the data employed. In this regard, rigorous testing, validation and access to high quality data are essential to ensuring high quality computational models.

Cross validation is a standard technique used for ML model performance evaluation and operates under the assumption that each datapoint in the dataset is independent from one another. However, this is not necessarily the case in biomedical data due to high sample/variable correlations, high- dimensionality, lurking variables, and technical biases. due to the presence of hidden duplicates. This problem has been termed as the 'doppelgänger effect', a phenomenon where unrelated/independently derived biological samples appear very similar to each other by chance. Doppelgänger effect is problematic as if left undetected, can exaggerate ML performance, resulting in misleading confidence and possibly wrong (not biologically significant) outcomes to be concluded. Therefore, it is crucial to identify the presence of doppelgängers in datasets between before model validation. It is also important to realise that doppelgängers are not necessarily functional, i.e. have the ability to confound ML outcomes. For example, certain doppelgängers may simply be a result of a data leakage between sample replicates, and thus may not exert the doppelgänger effect.

Unfortunately, there is an overall lack of consistency in checking for such chance similarities in training and test data sets. Moreover, data doppelgänger and its implication on ML are poorly documented and not well-understood. Though doppelgängers have been previously detected in biomedical data, majority of the identification methods used are not sufficiently robust or has poor generalizability.

2. Rationale

In response, Wang et al. (2022) conducted their study to prospectively evaluate whether data doppelgängers are present in a renal cell carcinoma proteomics dataset, and subsequently investigate the effects of data doppelgänger effect on ML and find ways to minimize the doppelgänger effect.

3. Methods & Results

The dataset was chosen as it has well-defined meta-data, which allowed distinct evaluation scenarios to be constructed. They include a variety of negative cases where data doppelgängers cannot exist (different classes), and positive cases in which doppelgängers are possible due to leakages (same patient and same class based on replicates). Main steps of their identification method are 1) batch correction, 2) calculating the Pairwise Pearson's correlation coefficient (PPCC) between samples of different datasets, 3) grouping sample pairs by similarities of their patient and class, 4) calculating and applying a cut-off, which is the maximum PPCC of any sample pair in the negative case, i.e., 'Different Patient Different class'. The PPCC distribution of the valid scenario, 'Same Class Different Patient', were compared against the negative and positive scenario. The PPCC data doppelgängers were then identified as valid sample pairs existing at the high end of the distribution, specifically with PPCC values above the cut-off. In sum, 26 PPCC data doppelgängers were identified from the renal cell carcinoma proteomics dataset. Another interesting observation was that in the strip plots, the PPCCs in the valid scenario exists as a continuum distribution without any visible breaks. This suggests that data doppelgängers can exists naturally as part of the dataset without appearing as extreme anomalies. Therefore, using outlier detection methods may not be sufficiently sensitive.

The PPCC data doppelgängers identified were also found to act as functional doppelgängers, exerting a direct inflationary effect on the accuracies of different trained ML models. This was achieved by first stratifying the test data into two different strata - PPCC data doppelgängers and non- PPCC data doppelgängers strata and then performing separate evaluations on the ML model performance. All ML models on the PPCC data doppelgängers strata consistently performed better as compared to on the non-PPCC data doppelgängers strata. Moreover, the number of PPCC data doppelgängers were found to be positively correlated to performance inflation, specifically demonstrating an additive inflationary effect. When all 8 doppelgängers were included in the validation set, the accuracy distribution was identical to the training-validation set with perfect leakage (i.e. the positive control in their experimental setup).

4. *Limitations*

Though the PPCC data doppelgängers method was able to identify functional data doppelgängers, there are a couple of limitations in this study. First, Wang et al. (2022) was not able to determine a robust method that directly removes data doppelgängers. It was discovered that the doppelgänger effect could easily be eliminated by constraining all identified PPCC data doppelgängers to either the training set or validation set. However, such an approach is not ideal given that it would lead to ML models that in the case of the former, is incapable of generalizing well due to inadequate knowledge. In the latter case, the models might produce extreme scenarios where the doppelgängers are all predicted correctly or wrongly. Moreover, the scope of their experimental setup is unfortunately limited to a dataset that comes from a single disease type and derived from a specific platform, i.e. mass spectrometry. It also does not account for other types of datasets such as integrated data analysis or meta-analysis. Moving forward, it would be interesting to evaluate the doppelgänger effect on a wider range of diseases and data types.

5. Doppelgänger effects in other types of biomedical data

The doppelgänger effect is a phenomenon that has also been detected in other types of biomedical data such as gene expression datasets. Cancer specimens are often re-used or shared in clinical genomic studies, which can lead to duplication of expression profiles in public databases. If left unchecked, this duplication leads to a doppelgänger effect which inflate the accuracy of genomic ML models when integrating data from different studies. To determine the presence of doppelgänger effect in cancer expression profiles, Waldron et al. (2016) evaluated different cancer microarray profiles and RNA sequencing expression profiles taken from The Cancer Genome Atlas (TCGA) using a PPCC outlier detection package, doppelgangR. In general, they compared dataset pairs and samples pairs to create an empirical distribution of batch-corrected pairwise correlations between biological replicates within a single dataset or between two different datasets (where different profiling technologies may be used). Duplicate expression profiles with unusually high pairwise correlation were then identified as outliers. As their dataset involved different cancer types as well as profiling technologies, pairwise PCC values were determined between either biological replicates within a dataset or between datasets when different profiling technologies were used. More than half of all studies evaluated were to found to contain doppelgängers. For example, for the breast cancer database, out of 1467 gene expression profiles studied, 59 samples were identified to be duplicates in two different published studies. Waldron et al. (2016) attributed this duplication to a group of samples shared between the datasets. Another finding was that the proportion of duplicates had a additive effect on the validation set hazard ratio, suggesting the presence of an inflationary effect. For example, a 30% duplication increased hazard ratio from 1.1 to 1.7. Overall, the doppelgangR approach was demonstrated to work when the cancer expression profiles are distinct enough. For example, for cancer types with low PCCs, they typically contain widespread genomic alterations which result in distinctive gene expression profiles that make it possible to distinguish and identify the doppelgängers.

6. Recommendations

To achieve good quality ML models, it is important to check for the doppelgänger effect. One recommendation is to perform cross-checks in the meta-data to ensure that sample between the training and test sets are truly dissimilar/not of high similarity. Even if the training and test sets are independently derived, a doppelgänger effect may be present which does not inform the true learning performance of the model despite having good validation outcomes. (Ho et al., 2020) As demonstrated by Wang et al. (2022), cross-checks enabled them to create positive and negative cases as reference points, which aided their identification of potential data doppelgängers. Another suggestion is to perform data stratification where data is first stratified into different similarities and then evaluating the model performance on each stratum separately. Strata that showed poor model performance can then be investigated, for example they can be a potential area of weakness for classifier and require need further improvement. It is also highly recommended to perform independent validation checks with multiple independently derived data sets. Internal validation approaches such as cross-validation cannot guarantee for quality of the ML model given the original input data may not be truly representative or properly prepared. In contrast, divergent validation helps us to

study the generalizability of the model in terms of real-world usage (if it is universally applicable) but also can be used to check for issues with the validation data. (Ho et al., 2020)

7. References

Ho, S. Y., Phua, K., Wong, L., & Bin Goh, W. W. (2020). Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. *Patterns (New York, N.Y.)*, *1*(8), 100129. https://doi.org/10.1016/j.patter.2020.100129

Levi Waldron, Markus Riester, Marcel Ramos, Giovanni Parmigiani, Michael Birrer, The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles, *JNCI: Journal of the National Cancer Institute*, Volume 108, Issue 11, November 2016, djw146. https://doi.org/10.1093/jnci/djw146

Wang, L. R., Wong, L., & Goh, W. W. B. (2022). How doppelgänger effects in biomedical data confound machine learning. *Drug discovery today*, *27*(3), 678–685. https://doi.org/10.1016/j.drudis.2021.10.017