Question 3

Report of *How doppelgänger effects in biomedical data confound machine learning*

1. Introduction

Machine learning has been increasingly adopted in biology in recent years and has profoundly influenced its development. For example, in drug discovery, machine learning increases the efficiency of drug discovery in a multitude of ways: machine learning models can shortlist better drug candidates faster, reducing time spent on discovery and testing, which have speeded up drug development.

Cross-validation techniques are commonly used to evaluate these machine learning models. However, the reliability of such validation methods can be affected by the presence of data doppelgängers. Data doppelgängers occur when independently derived data are very similar to each other, causing doppelgänger effects, which describe the situation when a machine learning model performs well on a validation set regardless of how it has been trained [1]. Doppelgänger effects are problematic as they could exaggerate the performance of the machine learning model on real-world data and potentially complicate model selection processes that are solely based on validation accuracy [1]. Hence, it is crucial to be aware of the presence of any doppelgängers before model validation and avoid them in the practice and development of machine learning models.

2. Abundance of doppelganger effects in biological data

In addition to the biological data mentioned in the original article, such as the single renal cell carcinoma (RCC) protein expression (proteomics) dataset, doppelganger effects also can be observed in RNA-Seq and microarray gene expression data. For example, the relevant literature demonstrates the doppelganger effects of the well-studied microarray gene expression data from the study of Belorka and Wong and the widely available RNA-Seq gene expression data from the Cancer Cell Line Encyclopedia (CCLE) project [1-3].

3. Doppelganger effects are not unique to biomedical data

Although doppelgänger effects have been observed in biomedical data, they are not unique to biomedical data. There are two key definitions related to doppelgänger effects – data doppelgängers and functional doppelgängers. Among them, data doppelgängers are sample pairs that exhibit very high mutual correlations or similarities, and functional doppelgängers are sample pairs that result in inflated machine learning performance when split across training and validation data, namely the machine learning will be accurate regardless of how it was trained [1]. They are responsible for doppelganger effects. Therefore, it is clear that the generation of doppelgänger effects is not an intrinsic issue with biomedical data, as other types of data are also equally likely to have data doppelgangers and functional

doppelgangers, which are not unique to biomedical data and can cause doppelgänger effects.

4. Identification of data doppelgängers and functional doppelgängers

Since data doppelgängers and functional doppelgängers are responsible for doppelganger effects, identifying them is of great help in checking for doppelganger effects.

Based on the original literature, the basic design of the pairwise Pearson's correlation coefficient (PPCC) as a quantitation measure is reasonable methodologically. The results of identifying PPCC data doppelgangers based on the PPCC distribution of the valid scenario against the negative and positive scenarios showed that the presence of PPCC data doppelgängers in both training and validation data inflates machine learning performance even if the features are randomly selected and it is consistently reproducible on different sets of training and validation data and on different machine learning models. Moreover, the more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance, pointing toward a dosage-based relationship between the number of PPCC data doppelgängers and the magnitude of the doppelgänger effect. Thus, PPCC can be used to identify potential functional doppelgangers from the constructed benchmark scenarios, and the PPCC data doppelgängers based on pairwise correlations act as functional doppelgängers, producing inflationary effects similar to data leakage.

Though the PPCC data doppelgänger identification method has been proven to be able to identify functional doppelgängers, there is still room for improvement. For example, in some cases, some doppelganger effects persist even where no PPCC data doppelgängers should exist between training and validation sets, which may suggest that functional doppelgängers still exist between the training and validation sets but are undetectable by PPCC. This could be the result of the shortcomings of Pearson's correlation and non-normality may possibly have reduced the effectiveness of Pearson's correlation in detecting data doppelgängers. Another possible hypothesis is Pearson's correlation's inability to capture non-linear relationships and therefore, it is unable to identify more complex functional doppelgängers that are linearly dissimilar but non-linearly associated. Based on this, future work focusing on incorporating other similarity metrics robust to non-normal data like Spearman's measure or metrics capable of detecting non-linear relationships like dCor into the data doppelgänger identification procedure may be helpful [1]. Another possible reason for poor functional doppelgänger recovery could be attributed to the presence of anomalously high "Different Class Different Patient" PPCC values, which would inflate the PPCC cut-off reducing the number of detected PPCC data doppelgängers. Therefore, perhaps future work could attempt to alter the definition of the PPCC cut-off to be more robust to outliers [1].

5. Methods to avoid doppelganger effects in the practice and development of machine learning models

Although there seems to be no accepted way to avoid the doppelganger effects, the author gives some useful recommendations to avoid doppelganger effects in the practice and development of machine learning models in the original article.

The first recommendation is to perform careful cross-checks using meta-data as a guide. With information from the meta-data, researchers can identify potential doppelgängers and assort them all into either training or validation sets, effectively preventing doppelgänger effects, and allowing a relatively more objective evaluation of machine learning performance.
The second recommendation is to perform data stratification. Instead of evaluating model performance on whole test data, researchers can stratify data into strata of different similarities (e.g., PPCC data doppelgängers and non-PPCC data doppelgängers, and evaluate model performance on each stratum separately). Assuming each stratum coincides with a known proportion of real-world population, researchers are still able to appreciate the real-world performance of the classifier by considering the real-world prevalence of a stratum when interpreting the performance at that stratum. Moreover, strata with poor model performance pinpoint gaps in the classifier, which can be further improved.
The third recommendation is to perform extremely robust independent validation checks involving as many data sets as possible (divergent validation)[4]. Although not a direct hedge against data doppelgängers, divergent validation techniques can inform on the objectivity of the classifier. It also informs on the generalizability of the model in terms of real-world usage despite the possible presence of data doppelgängers in the training set.
In addition to the above three suggestions, the authors also suggest some other possible approaches to avoid doppelganger effects. For example, constraining the PPCC data doppelgängers to either the training or validation set are suboptimal solutions. In addition, in studies in which the PPCC outlier detection package, doppelgangR was used for the identification of doppelgängers, PPCC data doppelgängers could be removed to mitigate their effects, although this approach does not work on small data sets with a high proportion of PPCC data doppelgängers, such as RCC[5-6].
In addition, other articles also put forward some suggestions. For example, experimenting with different data transformation techniques like Gene Fuzzy Score (GFS) [7] or feature generation may be possible approaches to neutralizing doppelganger effects in data sets. Identifying FDs before data splitting and avoiding assorting FDs across training and validation sets may help to mitigate doppelganger effects. Moreover, comparing the accuracies of fine-tuned ML models with randomly trained models for an unbiased assessment of its performance on unseen data during the model evaluation may also help [1].

Extra references

[1] Wang Li Rong and Choy Xin Yun and Goh Wilson Wen Bin. Doppelgänger spotting in biomedical gene expression data[J]. iScience, 2022, 25(8): 104788-104788.
[2] Belorkar Abha and Wong Limsoon. GFS: fuzzy preprocessing for effective gene expression analysis. [J]. BMC bioinformatics, 2016, 17(Suppl 17): 540.
[3] Ghandi Mahmoud et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. [J]. Nature, 2019, 569(7757): 503-508.
[4] Ho Sung Yang et al. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability[J]. Patterns, 2020, 1(8): 100129-.
[5] Lakiotaki Kleanthi 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, 2018: bay011.
[6] Ma Siyuan et al. Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. [J]. Genome biology, 2018, 19(1): 142.
[7] Belorkar Abha and Wong Limsoon. GFS: fuzzy preprocessing for effective gene expression analysis. [J]. BMC bioinformatics, 2016, 17(Suppl 17) : 540.