Answer to question 3

Zihang Yu

1. Introduction

Machine learning is very popular and advanced with many different applications cross industries in recent studies. In health and medical science, machine learning is extremely important. It has outstanding capabilities for disease prediction and is used to reveal patterns in medical data sources (Shailaja, Seetharamulu et al. 2018).

Although the excellent capabilities of machine learning are appealing, they bring additional considerations. One of the examples is that the doppelganger effect in biomedical data could confound machine learning. Data doppelgangers happen when two sets of independently obtained data are extremely similar to one another. Data doppelganger might have an impact on the reliability of cross-validation techniques when evaluating models, since the trained models might perform well regardless of the quality of training (Wang, Wong et al. 2021).

Therefore, this report aims to investigate if doppelganger effects are unique to biomedical data. Moreover, providing possible ways to avoid the doppelganger effects in machine learning for health and medical science.

2. Doppelganger effects in other areas

Doppelganger effects exist in gene sequencing data. For example, on the basis of operational genomic data, TargetFinder3 is a machine-learning technique that predicts enhancer-promoter interactions. When the samples are randomly divided into training

and test sets, the high level of resemblance between window characteristics of the positive samples is likely to overestimate the cross-validation results (Cao and Fullwood 2019).

Another example is the doppelganger effects in machine learning models for RNA secondary structure prediction. RNA families share a lot more similarities in secondary structures than in sequencing. As a result, splits might be made when, while taking sequence similarity into account, almost identical structures are present in both the training and testing data sets. Existing methods of splitting the training and testing set might not work properly due to the confounding similarities (Szikszai, Wise et al. 2022).

Additionally, doppelganger effects could also appear in the field of business and economy. A doppelganger brand image can undermine the perceived authenticity of an emotive branding narrative and, consequently, the identification value that the business offers to customers (Thompson, Rindfleisch et al. 2006).

3. Checking and avoiding for doppelganger effects

Previous research has been done to give recommendations and develop methods to identify and avoid doppelganger effects in the practice of machine learning models. There are mainly three recommendations for identifying the doppelganger effects. The first is cross-checking possible doppelgangers using meta-data as a guide, then grouping them all into either training or validation sets. Secondly, rather than assessing model performance on the entire test set of data, stratify the data. Finally, use as many data sets as possible in reliable independent validation checks (Wang, Wong et al. 2021).

Moreover, there is an R package called “doppelgangerIdentifier” which includes functions for doppelganger identification and verification, including pairwise Pearson’s

correlation coefficient data doppelgangers identification and data doppelgangers inflationary effect verification (Wang, Choy et al. 2022).

4. Declaration

The doppelganger effects might be observed in financial data, where two different investment strategies produce similar returns over time. They could also be observed in social science data, where two different interventions have similar impacts on a particular outcome. However, the writer cannot find academic literature to support those ideas. Future research could focus on the doppelganger effects in financial and social science data.

5. Reference

Cao, F. and M. J. Fullwood (2019). "Inflated performance measures in enhancer–promoter interaction-prediction methods." Nature genetics **51**(8): 1196-1198.

Shailaja, K., et al. (2018). Machine learning in healthcare: A review. 2018 Second international conference on electronics, communication and aerospace technology (ICECA), IEEE.

Szikszai, M., et al. (2022). "Deep learning models for RNA secondary structure prediction (probably) do not generalise across families." bioRxiv.

Thompson, C. J., et al. (2006). "Emotional branding and the strategic value of the doppelgänger brand image." Journal of marketing **70**(1): 50-64.

Wang, L. R., et al. (2022). "Doppelgänger spotting in biomedical gene expression data." Iscience **25**(8): 104788.

Wang, L. R., et al. (2021). "How doppelgänger effects in biomedical data confound machine learning." Drug Discovery Today.