How to understand the doppelgänger effects of data

Introduction

Machine learning (ML) models have been increasingly used in drug development, where the reliability validation of widely used cross-validation techniques is compromised by doppelgänger effects. Doppelgänger effects are fairly common in our test data, and it has a direct inflationary effect on ML accuracy. This, in turn, reduces the usefulness of ML for phenotype analysis and subsequent identification of potential drug leads.[1]

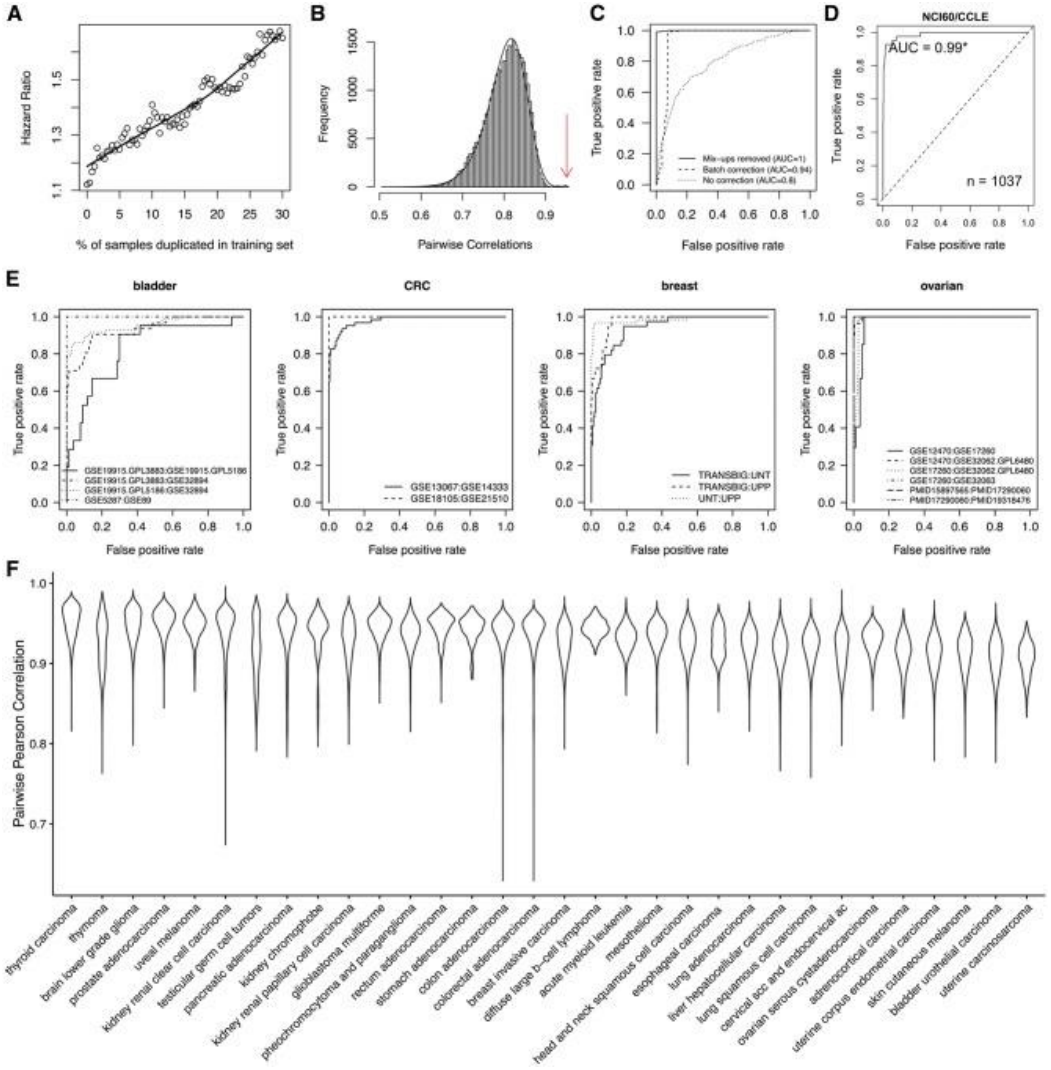

Figure 1.[2]

---

[1] Li Rong Wang, Limsoon Wong, Wilson Wen Bin Goh, How doppelgänger effects in biomedical data confound machine learning, Drug Discovery Today, Volume 27, Issue 3, 2022, Pages 678-685, ISSN 1359-6446, https://doi.org/10.1016/j.drudis.2021.10.017.

[2] Waldron L, Riester M, Ramos M, Parmigiani G, Birrer M. The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. J Natl Cancer Inst. 2016 Jul 5;108(11):djw146. doi: 10.1093/jnci/djw146. PMID: 27381624; PMCID: PMC5241903.

Doppelgänger effects in other fields

Machine learning (ML) models and the evaluation of cross-validation are not only used in biological fields such as drug development, but also many other fields, such as visual image processing and new material data processing industries. And according to research, the phenomenon of doppelgänger effects is strongly correlated with the way the data is validated and fed back, as well as the dataset itself. So for the biological domain, there is a strong similarity between the pattern of data set composition and the way algorithms are validated and the machine learning algorithms used in other domains, so it can be said that the data doppelgänger effects are prevalent.

Whole-genome analysis of cancer specimens is commonplace, and investigators frequently share or re-use specimens in later studies. Duplicate expression profiles in public databases will impact re-analysis if left undetected, a so-called “doppelgänger” effect.[2]

Figure 2.[1]

Analysis and Recommendations to doppelgänger effects

According to the study in Figure 2, the extent of this influence depends on two main factors: the similarity of functional doppelgängers and the proportion of functional doppelgängers in the validation set. Unfortunately, doppelgänger effects are not easy to resolve analytically. Therefore, to avoid performance inflation, it is important to check for potential doppelgängers in data before assortment in training and validation data.[1]

For dealing with the doppelgänger effects of data, two main approaches are proposed in this paper. Firstly, the training and validation data are processed to minimise the impact of doppelgänger effects. The second is to treat the duality validation as a separate model and algorithm optimization so that the training process has less impact on the doppelgänger effects of data.

Pre-processing the training and validation data is one of the main solutions to reduce the doppelgänger effects:

a) **Setting a threshold for correlation**

   When the correlation between two variables is ~|1|, retaining one or the other variable is irrelevant from a purely statistical point of view.

   The situation gets complicated when the correlation is lower than |1|, but still relevant, and a theoretical-based choice is impractical. In these cases, we need a criterion for determining whether a correlation is relevant or not. One strategy is to set a threshold on the correlation value.

b) **Pruning strategies**

   The key idea of pruning is dropping the most possible number of variables and retaining the greatest possible amount of information.

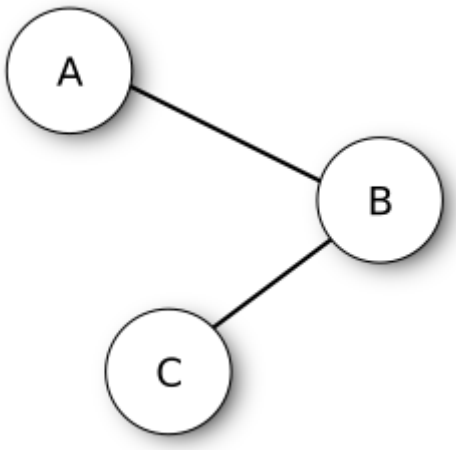

Figure 3[3]

For example, A, B, and C in Figure 3 have a strong relation. B has an over-threshold correlation both with A and C, while A and C have with each other an under-threshold correlation. A pruning approach can be dropping B, retaining the lowly

[3] Image from the author — released under license CC0

correlated A and C. However, a more parsimonious approach exists: retaining only B, dropping A and C. B can be considered the representative of the network and it can work as a spokesperson for the entire community.

For validation optimisation and algorithm optimisation, my recommendations are as follows:

a) **Performing extremely robust independent validation checks**
   Divergent validation on as much data as possible. Although not a direct hedge against data doppelgängers, divergent validation techniques can inform the objectivity of the classifier. It also informs on the generalizability of the model (in terms of realworld usage) despite the possible presence of data doppelgängers in the training set.[1]

b) **Prioritizing variables**
   Rank variables according to their centrality degree, intended as the strength of the connections of a node with other nodes. The most central nodes correspond to the most representative variables of the network. The ranking by centrality degree allows us to prioritize variables when choosing what to keep and what to drop.

Conclusion

The performance of ML models can be affected by strong doppelgänger effects, leading This article briefly summarises the commonality of data duality in different domains based on the information, as well as presents some insights into the elimination of doppelgänger effects. The main shortcomings of the article are the overall lack of experimental data support, all experimental data results from references, and the one-sidedness of the solutions to the doppelgänger effects, which needs to be studied in more depth.