A report on doppelgänger effects in biomedical data

WANG ZHOUCHI

Introduction

It is well established in machine learning that, the dataset should be spitted into distinct training, validation, and test sets, or the result will be biased and end up with a false impression of model accuracy. Doppelgänger effects describe a situation when samples exhibit chance similarities between training and validation data sets, causing misleading machine learning model performance [1, 2]. The author gives several instances of doppelgänger effects observed in bioinformatics, including the prediction of macromolecular interaction, protein functions, biological activities, etc. However, it is uncommon to check the similarities in training-validation pairs and the proposed methods for eliminating or minimizing similarity between test and training data are not robust enough.

Are doppelgänger effects unique to biomedical data?

Yes. The doppelgänger effects are unique to biomedical data, as they are prevalent in bio-related problems while uncommon in other areas. My understanding is as follows.

Individual elements or components can have very slight variances while having a large impact on the outcome of the analysis due to the complexity and variety of biomedical data, such as genomics, proteomics, and metabolomics. This could lead to distinct doppelgänger effects, in which identical patterns and trends in the data are observed that are distinct from non-biomedical datasets. Additionally, the differences caused by genetic and physical variability can also result in unique patterns in biomedical datasets that do not occur in other types of data. Doppelgänger data is indeed present in many datasets, however, in practical applications, chance similarities of samples rarely happen considering the very large sample size and relatively low probability of similar samples. Excluding resolvable conditions such as data leakage, doppelgängers surely exist but cannot lead to a functional doppelgänger effect (confounding ML outcomes), thus are not enough to constitute a particular concern in most non-biomedical areas.

The doppelgänger effects in biomedical data are intriguing in that although the training and test data sets are independently derived, they could still yield unreliable validation results. From my point of view, bio-related data has unique features and is very different from other data types. Modern big data often contains a large number of samples and a small number of features, while biomedical data always contains a limited number of samples but each sample contains a large volume of information, and most importantly, often has a high feature dimension. In biomedical data, there are random differences including temporal and spatial differences, as well as individual and population differences consistent with genetics,

physiology, and behavior. For example, tumor patients have differences in gene mutations at different sites, and their gene expression levels vary markedly from site to site and time to time, both of which are random and uncontrollable, and are difficult to eliminate.

This is where the dilemma of doppelgänger effects in biomedical data comes in. When analyzing other data types with a large number of samples and a small number of features or low data dimension (e. g. image recognition, traffic analysis, sales forecasting, etc.), machine learning can easily exclude differences between individuals and find commonalities between groups. When analyzing biomedical data based on a small number of samples with high-dimensional features, it is also easy for machine learning to find features that are relevant to the population. However, due to the problem of small sample sizes, the differences between individuals in the features found can easily be underestimated. Also, the temporal-spatial differences in the data due to sampling make the found features often poorly validated in new validation sets or the experiments of others.

How to avoid doppelgänger effects in machine learning models for biomedical data?

In the paper, the authors summarized several methods to identify data doppelgängers such as distributing samples in reduced-dimensional space and using the pairwise Pearson's correlation coefficient (PPCC), but removing data doppelgängers from data directly is quite elusive. The authors stated three recommendations to guard against doppelgänger effects:

- Perform careful cross-checks using meta-data as a guide
- Perform data stratification
- Perform extremely robust independent validation checks involving as many data sets as possible (divergent validation)

I think the main ways to avoid doppelgänger effects include these aspects:

(1) The biomedical data for training should be cleaned and pre-processed properly, including data normalization, class labeling, and format conversion. Clustering techniques can be used to identify and remove redundant or irrelevant attributes and reduce the number of input variables in the model. The core requirements are sufficient data and adequate sample size, checking for bias, heterogeneity, noise, and other confounding factors [3].
(2) For the complex features in biomedical data, techniques such as association rule mining can be applied to identify the most meaningful attributes, and many feature selection methods such as recursive feature elimination can also help. The biomedical significance of each feature must be clear. If necessary, manual work is needed to build new models or algorithms for different biomedical applications.
(3) An external test set can be used to evaluate the machine learning models [4]. Finally, cross-validation such as *k*-fold cross-validation and the Time-series-split method is important to minimize the potential doppelgänger effects.

References

[1] L.R. Wang, L. Wong, W.W.B. Goh, How doppelgänger effects in biomedical data confound machine learning, Drug Discovery Today 27(3) (2022) 678-685.
[2] L.R. Wang, X.Y. Choy, W.W.B. Goh, Doppelgänger spotting in biomedical gene expression data, iScience, 2022, p. 104788.
[3] S.Y. Ho, K. Phua, L. Wong, W.W. Bin Goh, Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability, Patterns 1(8) (2020) 100129.
[4] H.T. Xue, M. Stanley-Baker, A.W.K. Kong, H.L. Li, W.W.B. Goh, Data considerations for predictive modeling applied to the discovery of bioactive natural products, Drug Discovery Today 27(8) (2022) 2235-2243.