How doppelgänger effects in biomedical data confound machine learning

Abstract

Machine learning are enabling the study of biology and human health on an unprecedented scale and in multiple dimensions. These dimensions include numerous attributes describing the genome, epigenome, transcriptome, microbiome, phenotype, and lifestyle. However, no single data type can capture the complexity of all factors relevant to understanding a phenomenon such as disease. Duplicate expression profiles in public databases will impact re-analysis if left undetected, a so-called "doppelgänger" effect, and the reliability of validation methods that evaluates machine learning models can be affected by the presence of data doppelgängers. The main purpose of this report is to illustrate here the prevalence of functional doppelgänger under such biomedical data, the impact of data doppelgänger on ML, and methods to ameliorate the effects of doppelgänger, using a benchmark dataset of renal cell carcinoma.

Introduction

Machine learning models have been increasingly used in drug discovery to accelerate drug development. It is important that these trained models are properly tested to identify suitable drug candidates considering the expensive drug testing process. To evaluate the performance of a trained model, it is accepted that in ML, training and testing datasets should be derived independently. However, independently derived training and test sets may still yield unreliable validation results. Doppelgänger effects (DEs) occur when samples exhibit chance similarities, which inflate the performance of trained machine learning models when such similarities are split between the training and validation sets. This inflation effect can create misleading confidence in the deploy ability of the model. Thus, to date, there are no tools to identify the duo or standard practices to manage their confounding effects(Wang et al., 2022). Doppelgänger effects include data doppelgängers that are when samples appear similar across their measurements as well as functional doppelgängers that may confound ML outcomes. Therefore, there is an immediate need to investigate the property of data doppelgängers and to propose improved methods for doppelgängers identification so that understand the level of similarity between suspected functional doppelgängers and the acceptable proportion of functional doppelgängers in the validation set.

Prevalence of data doppelgängers in biological data

Data doppelgängers have been observed in modern bioinformatics and established fields of bioinformatics. Cao and Fullwood(Cao & Fullwood, 2019) revealed that the performance of chromatin interaction prediction systems has been overstated because of doppelgängers in assessment methodologies. In established protein function prediction, proteins with similar sequences are inferred to be descendants of the same ancestral protein and thus inherit the function of that ancestor. After further examination, however, we realized that this approach could not correctly predict the function of proteins with less similar sequences but similar functions, for example, enzymes with different overall sequences but similar active site residues(Friedberg, 2006). In addition, a similar example exists in quantitative structure–activity relationship (QSAR) models that are used to predict the biological activities of molecules from their structural properties. QSAR models assume that structurally similar molecules have similar activities. However, poorly trained models might still perform well on these molecules(Cherkasov et al., 2014), confounding model validation. Also, there isn't standard practice to eliminate or minimize similarity between test and training data before trained models evaluation.

Identification of data doppelgängers

Before validation, it is essential to be able to identify the data doppelgängers that exists between the training and validation sets. Ordination methods like principal component analysis or embedding methods, coupled with scatterplots, to see how samples are distributed in reduced-dimensional space can identify data doppelgängers. However, data doppelgängers is not necessarily distinguishable in the reduced dimensional space, thus, the method is not feasible. DupChecker(Sheng et al., 2014), identifies duplicate samples by comparing the MD5 fingerprints of their CEL files, also has the similar problems. Another method, the pairwise Pearson's correlation coefficient (PPCC) can captures relations between sample pairs of different data sets and anomalously high PPCC value indicates data doppelgängers. The fundamental of PPCC design is methodologically sound(Waldron et al., 2016). Therefore, it is used in this paper to identify potential functional doppelgängers from the benchmark scenarios that constructed from the renal cell carcinoma (RCC) proteomics data. Identifying PPCC data doppelgängers in RCC, there is a high proportion of PPCC data doppelgängers and PPCC distributions on the valid scenario exist as a wide continuum, without obvious breaks, which suggests that using outlier detection methods will not be sensitive enough. It also suggests that data doppelgängers exist naturally as part of the similarity spectrum between samples. PPCC values for same tissue pairs remain high overall, suggesting high correlations between samples, even if they come from different patients. These evaluations suggest that PPCC has meaningful discrimination value.

Confounding effects of PPCC data doppelgängers

After identifying PPCC data doppelgängers in RCC, the group explored their effects on validation accuracy across different randomly trained models. This would determine whether PPCC data doppelgängers act as functional doppelgängers, having an obvious inflationary effect on ML performances. They noted that PPCC data doppelgängers in both training and validation data inflates ML performance. Moreover, the more doppelgänger pairs represented in training and validation sets, the more inflated the ML performance. This indicates a dosage-based relationship between the number of PPCC data doppelgängers and the magnitude of the doppelgänger effect. Thus, PPCC data doppelgängers can act as functional doppelgängers, producing inflationary effects that are similar to data leakage. In addition, when all PPCC data doppelgängers are placed together in the training set, the doppelgänger effect is eliminated. This provides a possible way to avoid the doppelgänger effect.

Ameliorating data doppelgängers

How doppelgänger effects could be managed is challenging because many measures predicates on the existence of prior knowledge and good quality benchmarking data. First of all, the PPCC outlier detection package, doppelgangRwas used for the identification of doppelgängers(Lakiotaki et al., 2018), but this approach does not work on small data sets with a high proportion of PPCC data doppelgängers, such as RCC, because the removal of PPCC data doppelgängers would reduce the data to an unusable size. Second, the group intended to mitigate doppelgänger effect with data trimming by removing variables contributing strongly toward data doppelgängers effects. Nevertheless, there is no change in the inflationary effects of the PPCC data doppelgängers after the removal of correlated variables. This observation suggests the extreme complexity of the doppelgängers effect, as the reason for the high correlation between sample pairs cannot simply be explained by a subset of highly correlated variables. Therefore, it is time for us to look toward novel feature engineering and normalization approaches

Recommendations

Firstly, performing careful cross-checks using meta-data as a guide. The meta-data in RCC for constructing negative and positive cases allowed us to anticipate PPCC score ranges for scenarios in which doppelgängers cannot exist and where leakage exists. At the same time, when ML models are trained on data derived from biological sequences, researchers should ensure that training and test samples are not duplicates or samples of high similarity. Secondly, performing data stratification. Stratifying data into strata of different similarities and assuming each stratum coincides with a known proportion of real-world population, we 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. Last but not least, performing extremely robust independent validation checks involving as many data sets as possible(Ho et al., 2020). Future research could explore other methods of functional doppelgänger identification that do not rely heavily on meta-data. It is important to check the potential doppelgängers in the data before training and validating the data for classification in order to avoid performance inflation.

Reference

1. Cao, F., & Fullwood, M. J. (2019). Inflated performance measures in enhancer-promoter interaction-prediction methods. *Nat Genet*, *51*(8), 1196-1198.
2. Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., & Todeschini, R. (2014). QSAR modeling: where have you been? Where are you going to? *Journal of medicinal chemistry*, *57*(12), 4977-5010.
3. Friedberg, I. (2006). Automated protein function prediction--the genomic challenge. *Brief Bioinform*, *7*(3), 225-242.
4. 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 (N Y)*, *1*(8), 100129.
5. Lakiotaki, K., Vorniotakis, N., Tsagris, M., Georgakopoulos, G., & Tsamardinos, I. (2018). BioDataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology. *Database (Oxford)*, *2018*.
6. Sheng, Q., Shyr, Y., & Chen, X. (2014). DupChecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis. *BMC Bioinformatics*, *15*(1), 323.
7. Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016). The Doppelganger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. *J Natl Cancer Inst*, *108*(11).
8. Wang, L. R., Choy, X. Y., & Goh, W. W. B. (2022). Doppelganger spotting in biomedical gene expression data. *iScience*, *25*(8), 104788.