Data Doppelgängers

Ruining Cui 2022.12.20

Abstract

Machine learning (ML) models are increasingly being adopted for the processing of biomedical data, but the presence of data doppelgängers can mislead machine learning and cause model performance inflation. This doppelgängers effect occurs in a variety of fields, and it is essential that data is processed and identified to effectively reduce the damage caused by such effect.

Introduction

In the era of Big Data, individuals are generating data daily, which is collected and integrated for all aspects of life - monitoring of travel, analysis of diseases or prediction of personal preferences. This massive amount of information is captured and stored in various different databases, so it is easy to find data doppelgängers when integrating data from multiple shared databases. Data doppelgängers can occur in any field, but for biomedical science, publicly available human genomic data is often aggregated at a level where it cannot be uniquely identified in order to protect the privacy of the individual patient, leading to data doppelgängers being hard to identify. When data doppelgängers are independently derived and reflect extremely high similarity, the predicted model will exhibit good accuracy regardless of how it is trained, i.e. doppelgänger effects (DE). [1]

Data doppelgängers in different fields

Although the definition 'doppelgängers effects' was first coined in the field of biomedicine, and we have to admit that data doppelgängers in complex and unidentifiable medical data are more likely to be hidden from detection and successfully blend in with the databases of various models, it does not mean that doppelgängers effects are unique to biomedicine, because similar situations occur in all areas of life. Each domain has its own way of identifying these latent data

doppelgängers in order to avoid doppelgängers effects. For example, in the field of digital gazetteers, pairs of gazetteer records can be classified as data doppelgängers or non- data doppelgängers by using support vector machines or alternating decision trees with different combinations of feature vector similarity scores. [2]

Data doppelgänger and functional doppelgänger

In doppelgänger effects, there are two key definitions - data doppelgänger (DD) and functional doppelgänger (FD).[3] DD is a sample pair that exhibits very high correlation or similarity in measures of inter-sample relationships, which may lead to inflated ML performance (when acting as FD), but can also have no effect on ML training (non FD). FD is the culprit of inflated ML performance and "false" learning in actual machine training. When FD is widely found in the validation set, it can lead to accurate results regardless of how ML is trained. The fundamental measure to avoid the doppelgänger effects is to avoid the presence of FD in the dataset, although theoretically no matter what method is used, only the DD/FD ratio can be increased, while the possibility of non-DDs acting as FD can never be completely avoided.

Identification of data doppelgänger

In a critical model of biomedicine, several approaches are applied to identify data doppelgänger: 1. check the sample identifiers available in all datasets,[1] 2. batch correction, 3. batch balancing, 4. select appropriate metrics to calculate the correlation between each sample in one dataset and in the other dataset, 5. validate the obtained DD for FD before data splitting,[4] 6. validating the dataset containing the FD is processed, 7. using different data transformation techniques such as genetic fuzzy systems (GFS) or feature generation.

The biomedical field often utilises multiple datasets from different sources to improve statistical power and reduce uncertainty, but batch effects (BE) from this can confound statistical feature selection and mislead ML model training, making batch correction a necessity. And batch imbalance can potentially affect the identification of DDs, leading to a reduction or increase in the number of DDs identified, which allows

potential FD to escape detection. Taking the results of Pearson's correlation coefficient as an example, it can be seen from the graph below (Figure 1) that the number of PPCC dd's identified after batch balancing was performed changed significantly, with a decrease in the number of PPCC DDs in the DMD and ALL datasets and an increase in the Leukaemia dataset. Additional PPCC DDs were even identified after batch balancing in DMD and Leukaemia, suggesting that batch balancing may have contributed to the identification of PPCC DDs acting as FD.

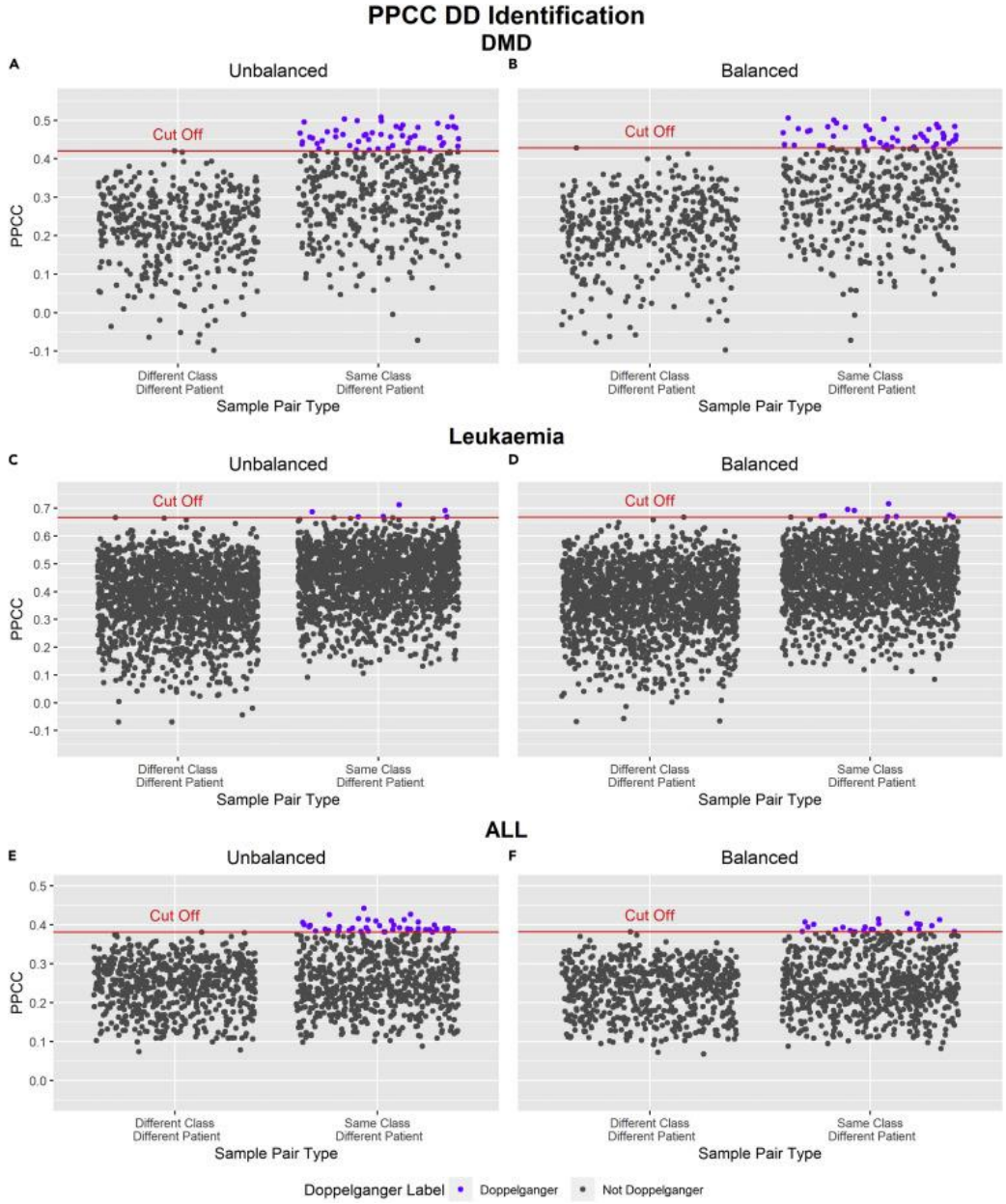

Figure 1 PPCC distributions sample pairs between DMD, leukemia, and ALL data sets with and without batch imbalance.[3]

In addition to the Pearson correlation coefficient, which is commonly used to calculate correlations between samples from different data sets, other correlations such as the Spearman's rank correlation coefficient and the Kendall's rank correlation coefficient are also useful for identifying DD, where the Pearson correlation coefficient is a measure of linearity and the Spearman's rank correlation coefficient and Kendall's rank correlation coefficient are measures of monotonic relationships between samples. The choice is generally based on the desired analytical sensitivity, precision and data distribution.

The validation of the obtained DDs as FD is crucial, as the manipulation of the FD in the dataset can significantly help to improve the performance of machine learning. The distribution of model accuracy over different training validation sets can be measured to determine whether the resulting DDs are FDs. Commonly used training validation sets are 'i Doppel' (i = 0,2,4,6,8,10, i refers to the number of samples of DD), 'i Pos Con' (i = 10,6,5, i refers to the number of duplicated samples from the training set) and 'Neg Con' (n = 10, p = 0.5, generated by 22 binomial distribution generated by accuracy). Based on the correlation between the number of DDs and the random model validation accuracy, we can roughly assume that the fraction of DDs detected as FD. When there is a positive correlation between the number of DD samples and the random model validation accuracy, most of the detected DDs are FDs. Based on the trend of increasing accuracy between different 'i Doppels', it can be inferred that the DDs added between the training validation sets are FDs.

The doppelgänger effects can also be effectively mitigated by dealing with the validated dataset; datasets with a small proportion of FD or a high volume of data can be mitigated by removing the FD, but this approach is not applicable to small datasets with a high proportion of DD mixes, as removing the DD would significantly reduce the volume of data and render it unusable. For the second case, avoiding classification of the FD in the training and validation sets and restricting it to the training or validation set is the next best solution. However, it is also prone to associated problems, such as when restricting the FD to the training set, it is easy to exclude a less similar sample due to the fixed size of the training set, causing the model not to

generalize well, while restricting the FD to the validation set, too extreme validation results can be obtained, with a detrimental inflationary effect on ML. Stratifying the data for the whole dataset and evaluating the model performance for each stratum separately, and evaluating the model as a whole based on the proportion of different strata in the real world is beneficial for us to understand how the model performs in the real world and to improve it for deficient models.[4]

Conclusion

Doppelgänger effects arise in all areas of data manipulation, since ML model performance is related to the cleanliness and accuracy of the data and requires independence between the validation and training sets, the false exaggeration of model accuracy by doppelgänger effects can reduce the reliability of machine learning models in real-world practice. Therefore, to reduce the doppelgänger effects, batch correction, batch balancing, searching for DDs, validating FDs and data stratification can be adopted to avoid machine learning performance inflation.

Reference

1. Waldron, L., Riester, M., Ramos, M., Parmigiani, G. & Birrer, M. The doppelganger effect: Hidden duplicates in databases of transcriptome profiles. J. Natl. Cancer Inst. 108, (2016).
2. Martins, B. A supervised machine learning approach for duplicate detection over gazetteer records. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol. 6631 LNCS (2011).
3. Wang, L. R., Choy, X. Y. & Goh, W. W. B. Doppelgänger spotting in biomedical gene expression data. iScience 25, 104788 (2022).
4. Wang, L. R., Wong, L. & Goh, W. W. B. How doppelgänger effects in biomedical data confound machine learning. Drug Discov. Today 27, (2022).