Liu Yu Xuan

Döppelganger effects in biomedical data

1. Introduction

In machine learning (ML), computer systems are conferred the ability to learn on their own and improve performance without direct instructions. Through the use of data and algorithms, machines simulate humans' capacity to recognise patterns, but objectively. This is especially helpful when the dataset to be analysed is too vast or complicated, or when there is a need to automate the process. Biological data usually possess these properties, and with its rapid generation, ML is becoming increasingly important in this field. Instead of relying on direct input commands, the machine perform tasks based on input data received. Training data is the initial part of data used to build the ML model, which repeatedly analyses the training data and modifies itself to achieve its intended function. Validation data is new and unevaluated data introduced to the ML model to serve as an initial assessment against unfamiliar data and to help with hyperparameter optimisation (Theobald, 2021).

Data döppelgangers (DDs) refer to data samples that appear very alike one another, and when the training and validation sets are too similar, a döppelganger effect (DE) is generated whereby the performance of the ML model would be exaggerated. This is because the model can perform well by simply memorising the data from the training set and is not being properly evaluated on its predictive abilities. DE is a concerning issue as it may lead to inflated performance of ML models on real-world data, or to the selection of inappropriate models (Wang *et al*., 2022b).

2. Döppelganger effects in biological data

DEs have been found in biological data in several instances. This section will cover some examples and discuss whether DEs are unique to biological data.

TargetFinder is a ML method that predicts interactions between enhancers and promoters based on the profile of transcription factors and histone modifications in the window regions between them. However, performance was found to be inflated, with interactions predicted at random parts in the window regions. This was due to extensive overlap between window regions of positive samples, and performance was subsequently lowered when training and test data were split based on individual chromosomes (Cao & Fullwood, 2019).

Gene signatures that are associated strongly with disease outcome are useful in indicating the gene expression related to that disease. However, published breast cancer signatures have been found to perform just as well as random ones when evaluated with an independent data set, and irrelevant non-breast cancer signatures also performed well. This suggests that many breast cancer signatures consist of proliferation genes which

are not exclusive to the disease. Random signatures are significant in multiple data sets, possibly due to similarities between breast cancer data sets despite them being independently derived. The random signatures themselves could also contain similar genes and thus were not completely independent from one another. Therefore, a data set could contain genes common to a few random signatures and all of them would be found to be significant there (Goh & Wong, 2019).

In the prediction of protein function, proteins with similar sequences are assumed to have similar functions. While this is true in many cases, a ML model trained this way will be unable to make accurate predictions for two proteins with different sequences but similar functions, such as for twilight zone homologues and enzymes similar only in their active sites (Wang *et al*., 2022a).

For ML models that take in protein sequence data as input, researchers try to make sure no data in the training set is too similar to data in the test set, at a threshold of about 25% similarity. While this approach can prevent many homologous pairs of proteins from being split between training and test sets, there are many homologous proteins with little sequence similarity and thus filtering by sequence alone will be insufficient in preventing DEs. According to the ML model, such proteins will have identical profiles. Possible solutions include finding sequences distantly related to training data by searching the test data with hidden Markov model comparison tools. When the input or output involves protein structure, structural classification such as Evolutionary Classification of protein Domains can be used to exclude homologous proteins (Greener *et al*., 2022). The prediction of protein-ligand interactions is affected by similar issues, whereby high similarity between training and test proteins was found to inflate performance of ML-based methods. After removal of training proteins that were too similar to test proteins, ML-based methods no longer outperformed conventional scoring methods (Li & Yang, 2017).

In drug discovery, the Quantitative structure-activity relationship (QSAR) model predicts the biological activities of molecules based on their structure, and thus assumes that structurally similar molecules have similar activities. An issue similar to protein structure prediction is encountered- while the assumption may hold true most of the time, the distribution of molecules with similar structures and activities into both the training and test data set might cause the model to be unable to perform well on molecules with similar structures but different activities. Only well-trained models can identify small variations in structure that lead to differences in activity (Wang *et al*., 2022a).

Recently, several ML models reported impressive results for RNA secondary structure prediction but only for intra-family predictions, whereby overlap between families in training and test data sets are expected. Meanwhile, inter-family predictions, whereby no such overlap is expected, are not addressed extensively. Considering that RNA structure is highly conserved within families, the prediction by such ML models do not generalise across families (Szikszai *et al*., 2022).

The above examples illustrate the prevalence of DEs in biomedical data, which commonly contain similar data points that are eventually distributed among training and validation sets. Data could be similar to one another due to pure coincidence or assumptions that neglects some aspects of the data (e.g., in protein function prediction, excluding homologous proteins based on sequence, and drug discovery). The ML model thus only performs well when presented with familiar data in the validation data set.

This phenomenon is not exclusive to biomedical data and could be observed in other fields. For example, the demand for a food item could be relatively stable for most parts of the year, and a ML model can accurately predict sales based on past data. A cultural celebration can lead to a temporary sharp increase in sales, but if knowledge of this celebration and its specific dates are lacking or omitted, the ML model will not be able to model this peak to accurately predict sales of this food during the celebration period of the following year. Therefore, DEs can be observed in other fields when DDs arise from analysing data with insufficient domain knowledge.

3. Identifying functional döppelgangers

DDs can be identified with the pairwise Pearson's correlation coefficient (PPCC), which was first applied to identify duplicate profiles in ovarian, breast, bladder, and colorectal cancer databases (Waldron *et al*., 2016). Although the study did not show that the DDs identified had functional effects of inflating ML performance, and that the supposed DDs were actually due to data leakage, the use of PPCC to quantitatively measure the relationship of samples between different data sets is logical.

PPCC was subsequently used to identify functional döppelgangers (FD) in two renal cell carcinoma (RCC) proteomics datasets (Wang *et al*., 2022a). Following batch correction of datasets, PPCC between all sample pairs were calculated. PPCC values were classified into 3 cases: (i) positive cases which represent data leakage, (ii) valid cases whereby sample pairs could be DDs since they consist of the same tissue type (normal or tumour) but are from different patients, and (iii) negative cases whereby sample pairs cannot be DDs since they consist of different tissue types. Sample pairs among the valid cases with PPCC values higher than the maximum PPCC value among the negative cases were identified as PPCC DDs. It was then found that half the samples were PPCC DDs with at least one other sample.

The confounding effects of these PPCC DDs were then investigated on 4 different types of randomly trained models: k-nearest neighbours (kNN), naïve bayes, decision tree, and logistic regression. It was found that ML performance improves with increasing representation of DDs in both training and validation sets, thus confirming that PPCC DDs act as FDs and cause effects similar to data leakage. This effect was more pronounced in kNN and naïve bayes models, compared to decision tree and logistic regression models.

The RCC dataset is a single data set, with one disease, and derived from mass spectrometry. To demonstrate the prevalence of DEs in other data types with different sizes, assay platform and more than one disease, PPCC DDs were also investigated in another study involving RNA-Seq gene expression data of Haematopoietic and Lymphoid Tissue-Lung tumour cell lines (Wang *et al*., 2022b). DDs were verified as FDs through randomly trained kNN models, during which a positive relationship between model accuracy and the number of DDs was observed.

4. Avoiding döppelganger effects

DDs can be reduced to avoid DEs. As discussed earlier, when data is assumed to always have certain properties or when there is a lack of domain knowledge limiting the analysis of available data, seemingly similar data will contribute to DEs when distributed across training and test sets. To avoid this, the ML model needs to include exceptions to its general assumptions (e.g., in protein structure prediction) or recognise obscure patterns in the data (e.g., in drug discovery).

Identified PPCC DDs can be removed to limit their inflationary effects on ML performance. However, this approach is not feasible for small datasets with high proportions of PPCC DDs, since the resulting data size would be too small and new issues such as overfitting, poor model performance, and inability to capture complex patterns in the data will be faced (Wang *et al*., 2022a).

Although it is difficult to eliminate DDs, several practices can be adopted to as precautions against DEs. Meta-data can be used to construct positive, valid and neutral cases which allows for identification of DDs through PPCC values. Identified DDs can then be all be put in either the training or validation set to prevent DEs. However, this is not the best solution as putting all DDs into a training set of fixed size excludes less similar samples and leads to the model lacking knowledge, while putting all DDs into the validation set causes all of them to be predicted either correctly or wrongly. Data can also be stratified into PPCC DDs and non-PPCC DDs to be evaluated separately. This allows for the gauge of model performance if each stratum corresponds to a known proportion in the population, and weaknesses of the classifier will be highlighted in strata with poor model performance. Divergent validation, which involves validation checks against other data sets, will suggest if the model is generalisable despite the training set containing DDs (Wang *et al*., 2022a).

5. Conclusion

DEs are common in biological data. Although the issue of DDs is known to the scientific community, it is usually still assumed that the training data is independent from validation data and there is no standard practice in reducing similarity between the two. DEs lead

to inflation of ML performance which reduces their usefulness. Therefore, it is crucial to verify if DDs are present before splitting data into training and validation sets.

6. References

Cao, F., & Fullwood, M. J. (2019). Inflated performance measures in enhancer-promoter interaction-prediction methods. Nature genetics, 51(8), 1196–1198. https://doi.org/10.1038/s41588-019-0434-7

Goh, W. W. B., & Wong, L. (2019). Turning straw into gold: building robustness into gene signature inference. Drug discovery today, 24(1), 31–36. https://doi.org/10.1016/j.drudis.2018.08.002

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. Nature reviews. Molecular cell biology, 23(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Li, Y., & Yang, J. (2017). Structural and Sequence Similarity Makes a Significant Impact on Machine-Learning-Based Scoring Functions for Protein-Ligand Interactions. Journal of chemical information and modeling, 57(4), 1007–1012. https://doi.org/10.1021/acs.jcim.7b00049

Szikszai, M., Wise, M., Datta, A., Ward, M., & Mathews, D. H. (2022). Deep learning models for RNA secondary structure prediction (probably) do not generalize across families. Bioinformatics (Oxford, England), 38(16), 3892–3899. https://doi.org/10.1093/bioinformatics/btac415

Theobald, O. (2021, January 1). Machine Learning for Absolute Beginners: A Plain English Introduction (Third Edition).

Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016). The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. Journal of the National Cancer Institute, 108(11), djw146. https://doi.org/10.1093/jnci/djw146

Wang, L. R., Choy, X. Y., & Goh, W. W. B. (2022b). Doppelgänger spotting in biomedical gene expression data. iScience, 25(8), 104788. https://doi.org/10.1016/j.isci.2022.104788

Wang, L. R., Wong, L., & Goh, W. W. B. (2022a). How doppelgänger effects in biomedical data confound machine learning. Drug discovery today, 27(3), 678–685. https://doi.org/10.1016/j.drudis.2021.10.017