1. Executive summary

The main purpose of this report is discussing several key topics about doppelganger effects. The report will focus on the definition and introduction of doppelganger effects, the problems doppelganger effects may cause, where and when doppelganger effects exist, and how to solve or avoid doppelganger effects.

2. Introduction

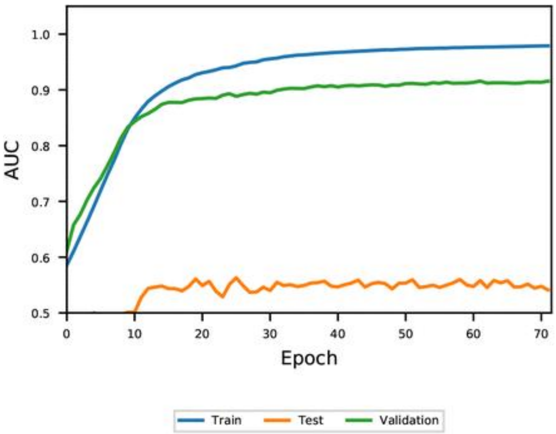
With the advent of machine learning and artificial intelligence era, machine learning methods have been increasingly used in the biomedical domain, including drug discovery, drug identification, etc. So machine learning models are very important parts in dealing with the biomedical data and the quality of the models after training process determines the performance of using models to do downstream tasks. However, although cross validation techniques have been used in model evaluation, some critical problems still need to be solved[1].

Doppelganger effects are very common problems which have been researched by researchers currently. Doppelganger effects exist when independently derived data are very similar to each other. Since they have almost same features, machine learning models are more likely to perform well without considering their training process. In order to avoid these effects, many studies have been conducted and there have been some methods to detect data doppelgangers and some solutions to mitigate the effects to some extent.

3. Discussion

3.1 Abundance of data doppelgangers

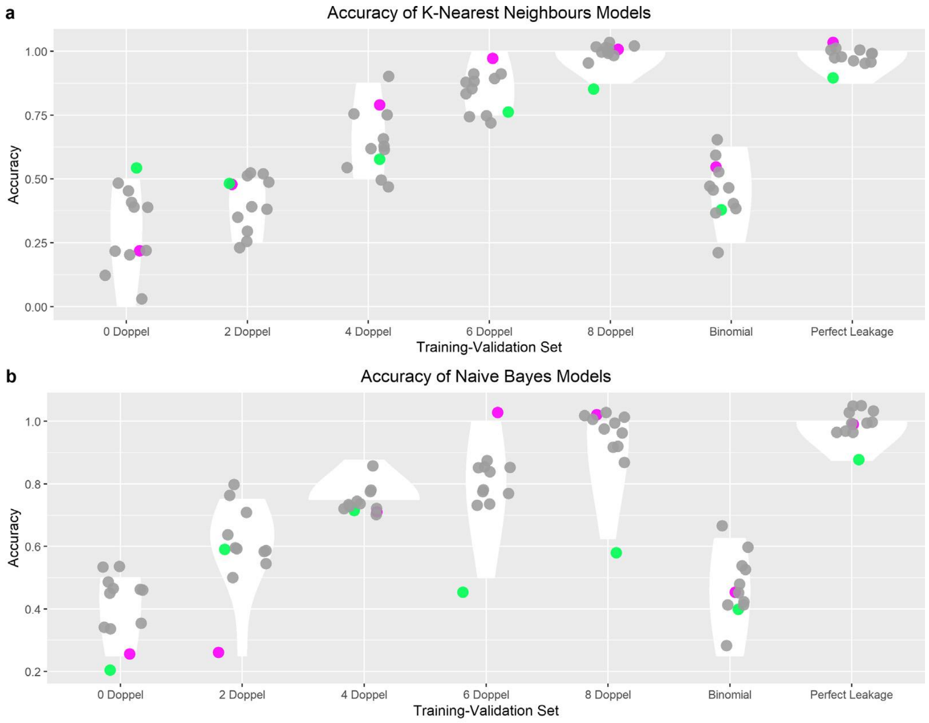
The presence of data has been observed in bioinformatics, such as the chromatin interaction prediction[2], protein function prediction[3], etc. Also, Doppelganger effects can be found in many interesting data types. Several independent studies have noted the presence of confounding similarities, including chromosomes, RNA families, or shared ancestry[4], between training and validation sets resulting in overinflated performance. Specifically, in one past studies, dataset about RNA-seq has concluded a quantity of data doppelgangers[5]. Since structure is so highly conserved amongst RNA families, current consideration of similarity measures is not enough. The model can achieve high Area Under Curve(AUC) in the intra-family case but fail to generalise whe it came to the inter-family case, which is shown below(cited from the paper[5]):

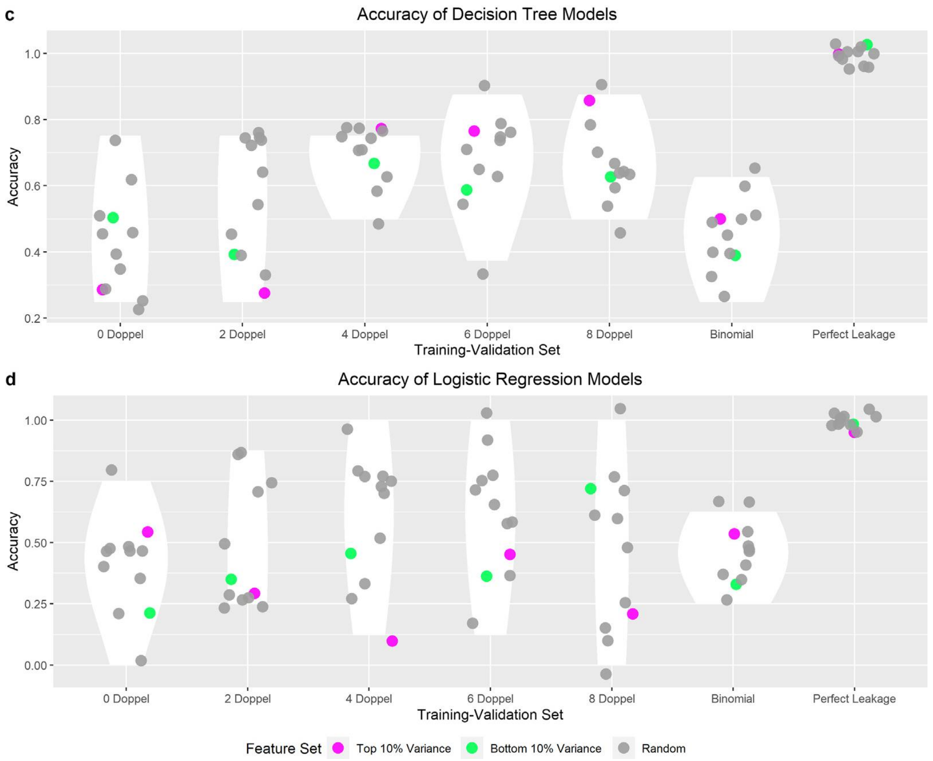


Besides, doppelganger effects are very common phenomonon in other areas. They are not unique to biomedical data. Doppelganger effects exist comprehensively in computer vision, natural language processing, etc. In domain of computer vision, it is not difficult for people to find pictures with same backgrounds. In many classification projects, if same-background pictures or duplicates are distributed into training dataset and validation dataset and are labelled in the same case by chance, the model will be more likely to classify the images successfully but not detect the real features during the training process, which will give a high accuracy on validation datast and a low accuracy on model generalizability. In domain of natural language processing, it is not rare for different texts that conclude some same sentence structure or preposition, which are very similar with each other. If these texts are distributed in the training and validation dataset, models are less likely to get meaningful features to do text classification generally. All in all, doppelganger effects are not unique to biomedical data.

3.2 Problems caused by data doppelgangers

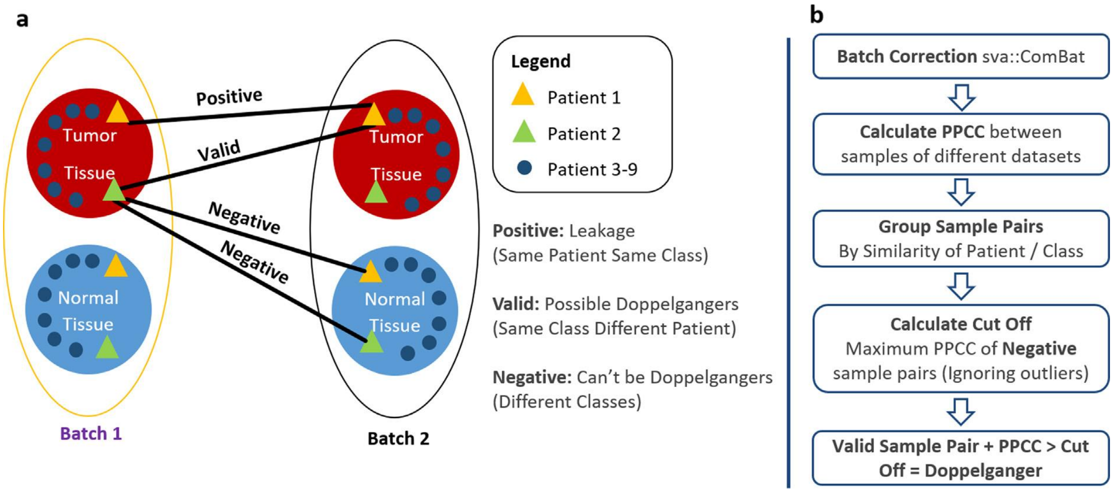
Since the data doppelgangers are very common and cause training troubles, a clear understanding is necessary. As we know, validation dataset serves as a small test dataset to evaluate the model after each training epoch. So if the validation performance is overinflated, researchers or early stoppers may stop training the model in order to prevent the models from overfitting. However, at that time, models may be underfitting or not be trained completely, which implies a much worse performance on test dataset than validation dataset. According to the past studies[1], data doppelgangers have different negative impact on different machine learning models. From a quantitative angle, with the increasing of data doppelgangers, models of K-Nearest Neighbours and Naive Bayes are harder hit by data doppelgangers than models of Decision Tree and Logistic Regression. We can find that if there are 8 doppelgangers in the training dataset and validation dataset, the performance is as same as the outcome of data perfect leakage happening, which is shown in the figure below(cited from the paper[1]):





3.3 Detection and Solutions to doppelganger effects

Since data doppelgangers will cause a series of problems, data doppelgangers should be identified before model validation. According to the past research[1], principal component analysis and embedding methods are not very useful. Early identifier Dupechecker also has some disadvantages. However, pairwise Pearson's correlation coefficient(PPCC) shows a big potential to detect the data doppelganger. The procedure to calculate PPCC of datasets is demonstrated in the figures from past research[1].


a
Tumor
Tissue
Normal
Tissue
Batch 1
Tumor
Tissue
Normal
Tissue
Batch 2
Positive
Valid
Negative
Negative
Legend
Patient 1
Patient 2
Patient 3-9
Positive: Leakage
(Same Patient Same Class)
Valid: Possible Doppelgangers
(Same Class Different Patient)
Negative: Can't be Doppelgangers
(Different Classes)
b
Batch Correction sva::ComBat
Calculate PPCC between samples of different datasets
Group Sample Pairs By Similarity of Patient / Class
Calculate Cut Off Maximum PPCC of Negative sample pairs (Ignoring outliers)
Valid Sample Pair + PPCC > Cut Off = Doppelganger

The presence of PPCC data doppelganger in both training and validation data inflates machine learning models. All models show better performance on PPCC data doppelgangers than non-PPCC data doppelgangers without considering generalizability. But unfortunately, currently researchers do not have a useful method to completely solve the problems caused by the data doppelgangers.

Besides Pearson’s correlation coefficient, the Spearman Rank correlation coefficient can also be used to identify data doppelgangers[4]. It is calculated by ranking the data within two different variables and computing the Pearson correlation coefficient on the ranks for the two variables. Spearman's correlation ranges in value from -1 to 1, with values near 1 indicating similarity in ranks for the two variables and values near -1 indicating ranks are dissimilar for the two variables. Spearman's correlation will be used to assess agreement in ranks between internal and external validation indices. Since it is based on ranks, it is less sensitive to outliers than the Pearson correlation coefficient and can also measure the strength of any monotonic relationship[6]. So it is useful to get relationships between two sets and find data doppelgangers from the training and validation dataset.

In order to avoid doppelganger effects in the practice of machine learning models, researchers can use three effective methods, mentioned in the past research[1], to mitigate. First, carefully checking the meta-data and creating the training dataset and validation dataset based on guidance from meta-data is one good direction to ameliorate these effects. Through diving into meta-data, researchers can get the structure and inner distribution of original dataset, so that they can control the PPCC score and assort all potential doppelganger data to either training dataset or validation dataset, ensuring training samples and validation samples that are not similar with each other. Second, researchers can stratify data into strata of different similarities. Evaluating model performance on each stratum separately is a good method but reaerchers also need to pay attention to ensuring that every stratum has enough samples for model training. So researchers may need to combine sample groups related several similarities together to create the training dataset and randomly pick out other samples to form the validation dataset. In this way, it is less likely to get a pair of training and validation datasets which include data doppelgangers effects and researchers will be able to find out the weakness easily because they can test their models on each stratum after training process. Third, creating a completely divergent validation dataset would be a indirect way for us to ameliorate the doppelganger effects. Obviously, if validation dataset is robust enough, the accuracy on validation dataset symbolizes the ability of model generalizability.

From all the discussion about how to avoid doppelganger effects during the practice of models, it is natural to come up with ideas that prevent doppelganger effects from happening in the development of machine learning models. On the one hand, it is effective to design a doppelganger detector adding in front of each machine learning model. This detector can be based on PPCC identification method illustrated above. So if the detector shows that the likelihood of doppelganger effects is very high, detector will inform the researchers to adjust the dataset. On the other hand, researchers can devise a complicated training procedure to mitigate this effect. They can design a sampler which is responsible for getting the training and validation datasets that conclude few PPCC data doppelgangers in each training epoch. Undoubtedly, samplers and detectors can be added into machine learning models cooperately so that machine learning models can be inputted datasets without data doppelgangers.

Except from these methods discussed above, since there are relatively less samples in the validation set, it is pratical to observe the subset of validation set which can be predicted well after fewer epochs. If the samples can be predicted really well only through one or two epochs training process, they are potential data doppelgangers in the validation set because the training process has not completed obviously. Then it is meaningful to use other models to predict them and check them as the next steps which has been mentioned in the paper[1].

Additionally, some researchers have created some useful packages in R to detect doppelganger[4]. The doppelgangerIdentifier R package allows users to easily identify PPCC DDs between and within data sets and verify the impacts of these detected PPCC DDs on ML model validation accuracy. And it provides a few functions for computation and visualization. In the future, doppelganger effects will be detected easier and solved faster.

4. Conclusion

From all the factors discussed above, doppelganger effects have wrecked havoc on model training, and these effects can not be solved easily. Currently, most of solutions to them are based on meta-data and coefficient matrices. Researchers need to check data doppelgangers carefully before model validation and adjust training or validation datasets. Future research on doppelganger effects is still needed.

5. Reference

[1] Wang LR, Wong L, Goh WWB. How doppelgänger effects in biomedical data confound machine learning. Drug Discov Today. 2022 Mar;27(3):678-685. doi: 10.1016/j.drudis.2021.10.017. Epub 2021 Oct 28. PMID: 34743902.

[2] F. Cao, M.J. Fullwood, Inflated performance measures in enhancer – promoter interaction-prediction methods, Nat Genet 51 (2019) 1196 – 1198.

[3] W.W.B. Goh, L. Wong, Turning straw into gold: building robustness into gene signature inference, Drug Discov Today 24 (2019) 31 – 36.

[4] Li Rong Wang, Xin Yun Choy, Wilson Wen Bin Goh,Doppelgänger spotting in biomedical gene expression data, iScience, Volume 25, Issue 8, 2022, 104788, ISSN 2589-0042

[5] Marcell Szikszai, Michael Wise, Amitava Datta, Max Ward, David H. Mathews Deep learning models for RNA secondary structure prediction (probably) do not generalise across families, bioRxiv 2022.03.21.485135;

[6] Khalid K. Al-jabery, Tayo Obafemi-Ajayi, Gayla R. Olbricht, Donald C. Wunsch II,7 - Evaluation of cluster validation metrics,Editor(s): Khalid K. Al-jabery, Tayo Obafemi-Ajayi, Gayla R. Olbricht, Donald C. Wunsch II, Computational Learning Approaches to Data Analytics in Biomedical Applications, Academic Press, 2020, Pages 189-208, ISBN 9780128144824