Report of question 3

Introduction

Machine learning has made great influence on drug development. However, unreliable validation results have been raised as an important problem in machine learning tasks. Due to highly similarities in biomedical data, such as protein structure, gene sequencing, drug molecular, etc., normal validation processes are hard to identify poorly trained models, thus there is a common challenge of ensuring robustness of the train models of biomedical data. This phenomenon is called Doppelganger effects. However, detection methods for such effect are still not effective enough since it remains uncharacterized [1]. Thus, it is important to find ways to identify the functional data doppelgangers which actually confounding the machine learning results and the leakage which has less effect on the training processes.

Uniqueness in biomedical data

Doppelganger effects can happen in situations where the data is inevitably highly similar. This can happen in many fields, besides biomedical data. One example can bird identification. Many different bird species are similar in appearance, even for professional birdwatchers, for instance, sparrows and a kind of small brown passerine birds called “little brown jobs”, especially for female birds with less colour compared to males [2]. To distinguish these birds, researchers have to take considered into many non-appearance factors, such as bird habits and sound wavelengths. Additionally, the task became even more challengeable where need to distinguish different individuals in same species.

How to avoid

In the case of health and medical science field, doppelganger effect can be very common. As mentioned in Wang et al. [1] paper, in protein function prediction and drug discovery cases, due to high similarity in protein sequences and molecule structures, it is very common to have train and validation datasets with similar patterns, thus the validation is less convincible.

There are some ways can possibly avoid or minimise such problems. Firstly, the data acquisition and generation process should be done with care. One possible way is to build up multi-centre databases. In different medical centres, medical devices, patients, caregivers can have many differences, if the data can be shared or even be public, the gathered datasets can be more robust due to more independently validation sets can be generated, and thus show the objectivity of the classifier. Another way to eliminate such problem is that, when generating datasets, it is important to ensure using different patients for training and test datasets. Individual differences can sometime be very large, in cases of gender, weights, heights, ages. If same patients’ data exist in both training and testing sets, the validation result will be seemly good, but the model may be poorly trained. In general, when acquiring or generating data, it is essential to keep those relatively highly correlated data in the same dataset, either training or validation sets. Consequently, training set and validation set can be less similar. In this case, the key point, where the small variation causing real biological activity, can be featured during training, such as the active site residual of proteins.

Besides of data collection, data quality measurements should be applied to help avoid doppelgangers, which means identify doppelganger effect first, before training and validation. Dimensional correlation is an important criterion showing the quality of data. Each feature should have relatively high correlation to the class or category, but uncorrelated with each other. If two dimensions has obvious positive or negative correlation, one of the dimensions should be considered abundant, and thus not useful or even harmful for data analysis. Therefore, correlation-based feature selection (CFS) is a good way to eliminate such problem. To examine the inter-correlation between different attributes, it can be evaluated by estimating the ability of one attribute predicting another. This can be achieved by decision trees. With CFS methods applied, it can produce a data subset containing major features predictive of the class and a relatively low level of redundancy. Consequently, increasing the quality of the dataset. Common CFS techniques including Pearson's correlation coefficient and ranks of nonlinear methods [3]. According to Wang et al. [1], among many previous works, Pairwise Pearson's correlation coefficient (PPCC) is a reasonable methodology that can quantitatively measure correlation between samples. It compared a pair sample from different data sets (batch), identify those cases of same class tissue in different patient as possible doppelgangers. However, although PPCC has good performance on finding redundant sample, it is not able to value the original one, thus, when cases with different importance level, it is likely to abundant the essential one. More importantly, PPCC cannot justify the ability of confounding machine learning tasks of the PPCC data doppelgangers [1].

However, existing data quality measurements in cases of doppelganger effects are not effective enough. Other methods for evaluate model performances are also required to show the quality of training. Wang et al. [1] recommended data stratification, which is to split data into more defined strata according to a set of criteria based on prior knowledge. For example, stratify data into doppelgangers and non-doppelgangers based on PPCC method, and evaluate them separately. Another solution from Wang et al. [1] was that perform cross-checks using meta-data as a guide. In this case, researchers are able to anticipate PPCC scores in all negative and leakage cases, so that the measurements can focus more on the real data doppelgangers samples that from same class but different cases. Therefore, instead of deleting possible doppelgangers directly, putting those data into either training or validation sets, so that they cannot existing in both sets.

Other data types

In image reconstruction problem, doppelganger effect also exists. To increase data size, data augmentation is a common method. For example, with 4 patients imaging data only, data simulation can perform in different starting acquisition angle, and different field of view, so that large amount of data can be generated. However, data from same patients can still be very similar to each other. In this case, it is important to measure the correlation between cases, and make those with high correlation into same set (training or validation). As mentioned above, patients for validation sets should be different to those used in the training set.

It also happens a lot in genome analysis. Waldron et al. [4] stated that, in whole-genome analysis of cancer specimens, it is common to use well-published later database, if did not detect duplicate expression profiles in public databases, the re-analysis quality would be impacted, thus leading to doppelganger effects. In this case, they proposed a doppelganger checking Bioconductor package from a huge number of studies.

Propose on avoiding method

Here I propose using meta learning to solve doppelganger problem. Meta-learning has been applied on small sample learning problem. In Wang et al. study, they stated that removing data doppelgangers from data directly would make the dataset unusable. However, meta-learning increases the possibility of making "unusable" dataset usable. In meta-learning, it uses support set and query set, which has similar function as training and testing set, however, if the support set has N classes items, and each classes has K shots, when $K < 10$, it can be considered as small sample [5]. Instead of training a single task on the model, the model should be trained on different tasks, and evaluation on different tasks. For example, in protein function prediction task, the model should be trained predicting different functions using the datasets with the same size of the support set. In this case, when support is very small, it can still be trained well. In another word, the model instead learns mapping relationship between "protein" to "label", it learns the mapping relationship between the "support set" to "c(.)", which is the mapping of "query set" and the "label".

In conclusion, meta learning provides the possibility of high training quality for small sample data by generating multiple subsets from original data, so that doppelganger effect can be minimized using very small database, instead of large but less effectiveness database.

Reference

[1] L. R. Wang, L. Wong, and W. W. bin Goh, "How doppelgänger effects in biomedical data confound machine learning," *Drug Discovery Today*, Oct. 2021, doi: 10.1016/J.DRUDIS.2021.10.017.

[2] E. Stokstad, "Not even scientists can tell these birds apart. But now, computers can," *Science*, American Association for the Advancement of Science (AAAS), Jul. 28, 2020. doi: 10.1126/SCIENCE.ABE0041.

[3] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," Hamilton, NewZealand, 1999.

[4] L. Waldron, M. Riester, M. Ramos, G. Parmigiani, and M. Birrer, "The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles," *JNCI Journal of the National Cancer Institute*, vol. 108, no. 11, Nov. 2016, doi: 10.1093/JNCI/DJW146.

[5] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," *34th International Conference on Machine Learning, ICML 2017*, vol. 3, pp. 1856–1868, Mar. 2017, Accessed: Feb. 02, 2022. [Online]. Available: https://arxiv.org/abs/1703.03400v3