A Review of the Ubiquity of Data Doppelgängers and Its Effect on Medical Data Analysis Based on Machine Learning

1. Abstract

Machine learning is widely used in analyzing large amount of data generated from biological or medical process. Generally, the evaluation of machine learning models depends on the validation set independent of the training set to verify the performance of the model. However, recent studies suggest that data doppelgangers in the validation set may lead to overestimation of ML model performance. This report will summarize the data doppelganger issues raised by relevant researches on human proteome analysis (Wang et al., 2021) and genome analysis (Waldron et al., 2016) (Wang et al., 2022). The widely used test methods for data doppelgangers are correlation analysis, such as Pearson correlation analysis. It will also discuss the possible effect of this phenomenon and propose some approaches to check and avoid the negative impact of data doppelgangers.

2. Background Introduction

In the field of biomedical data, machine learning is a widely used and proven effective data analysis method. Traditional machine learning methods include decision tree, MLP, SVM, KNN and other algorithms. The critical principle is training the model through statistical fitting of the primary extracted features to obtain the model that can predict the results of samples obtained in the real world. Machine learning can be used not only for analysis of biological structure and signal, but also for guidance and evaluation of medical programs and drugs (Garg & Mago, 2021). In a machine learning project, the evaluation of the model depends on training the model on the training set and testing its performance on the validation set (leave-out or k-fold validation). In this process, considering the ubiquity of overfitting in machine learning model, researchers have to ensure the generalization ability of the model to real-world samples. Thus, the data in the validation set must be independent of the training set. However, recent research suggests that there is a high proportion of doppelgangers data, which occur when samples in the validation set presents a high similarity to training samples, in biomedical data. For example, Waldron et al. (2016) and Wang et al. (2022) reported data doppelgangers is ubiquitous in multiple genome databases, and Wang et al. (2021) found out doppelgangers effect in renal cell carcinoma (RCC) proteomics dataset. Doppelgangers effect has been considered to have similar impact as data leakage and lead to overestimation of the model performance (Wang et al., 2021). These studies also put forward methods for quantitative identification of doppelgangers data (PPCC etc.), and proposed possible ways to avoid doppelgangers effect. It is worth noting that data doppelgangers can come from the structural

consistency of the same sample source, or from similar samples occurring in random sampling. The former can be easily solved by excluding samples from the same subject, while the latter needs further research. This report will mainly discusses the latter data doppelgangers.

3. Study methods and results on data doppelgangers

3.1 Pair Person Correlation Coefficient (PPCC)

Person correlation coefficient (PCC) is an algorithm to represent the linear correlation between two vectors. The general formula of PCC for a pair of vector X and Y is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$$where\ cov\ is\ is\ the\ covariance, \sigma\ is\ the\ standard\ deviation$$

The value of PCC is between - 1 and 1. An absolute value close to 1 represents a stronger the linear correlation between the two groups of variables. Physically, a closer PCC is to 1 means a smaller angle between the two vectors.

In machine learning dataset, a sample is usually described as a vector with fixed dimension. For example, in RCC dataset, a sample is represented as a vector with length of 12 (Wang et al., 2021). Therefore, it makes sense to calculate the Person correlation coefficient between two samples, which is called pairwise Pearson's correlation coefficient (PPCC).

3.2 Data Doppelgangers in Renal Cell Carcinoma (RCC) Proteomics Database

Wang et al (2021) used PPCC as an indicator to identify doppelganger data in the RCC database, and reported a high proportion of doppelganger data. This study grouped 36 samples by class and source. Sample pairs from the same patient are excluded to avoid structural data leakage. By calculating the PPCC of sample pairs from the same and different classes, Wang et al. found that sample pairs from the same class have a higher PPCC. When the PPCC value of the sample pairs from the same class exceeds the maximum PPCC value from different class of sample pairs (threshold), that sample pairs are defined as doppelganger data (purple points in Fig 1.a). Of the 36 samples in total, 18 were found to form data doppelganger with at least one other sample.

Wang et al. also reported significant confounding effect on machine learning model. The above data set is divided into training set and validation set (N=8). The training set is fitted with multiple machine learning algorithms such as KNN, and the classification accuracy obtained on the validation set is recorded and visualized in Fig 1.b. The proportion of doppelganger samples in the validation set data is the primary

variable. The results show that when the number of doppelganger data in the validation set increases from 0 to 8, the performance of the model in the validation set will significantly improve. Especially, if all samples are doppelganger, the accuracy rate can be close to 100%. However, when excluding all doppelganger data, the accuracy rate will be less than 50%, which is close to random guess. This result demonstrates that the existence of doppelganger effect is likely to cause overestimation of machine learning model performance. In the case of data leakage (same data in the training set and validation set), the performance of the model in the validation set is similar to that of the complete doppelganger data. Wang et al. concluded that doppelganger effect may possess a similar impact as the training data leakage.

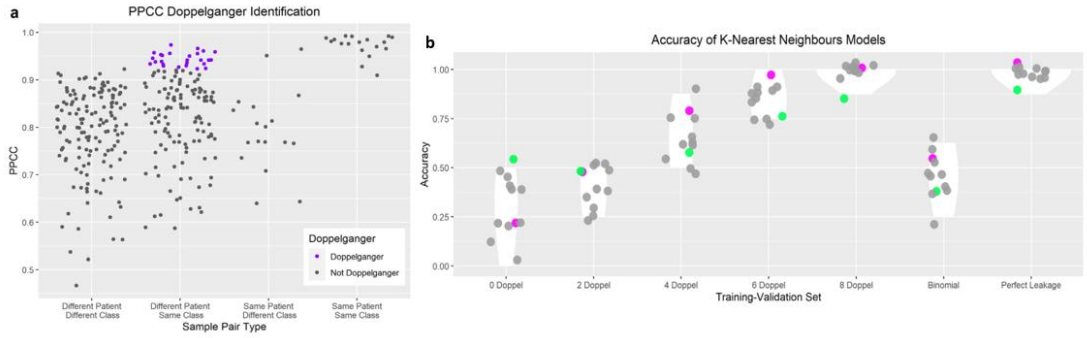

Figure 1(a). Distribution of PPCCs across different sample pairs. (b). The prediction performance of KNN models with different numbers of data doppelgängers in the validation set. (Wang et al. 2021)

3.3 Data Doppelgangers in Human Genome Database

In addition to the research on the doppelganger effect of proteomics, the doppelganger data in human genome research is also widely reported (Waldron et al., 2016) (Wang et al., 2022). These studies use similar PPCC threshold judgment method. Waldron et al. reported that in the of 17 groups RNA database of cancer cell, doppelganger data was found in 13 groups. Particularly, 569 doppelganger samples were found in 754 samples in GSE14333 and GSE17538 dataset.

Wang et al. (2022) also reported doppelganger effect of lymph_lung and large_upper dataset in RNA-Seq database. However, when validating the machine learning model on large_upper dataset, no significant improvement in the performance of the model was found when the doppelganger data increased. Therefore, further research and discussion are needed on the actual effect of the doppelganger effect on the machine learning model.

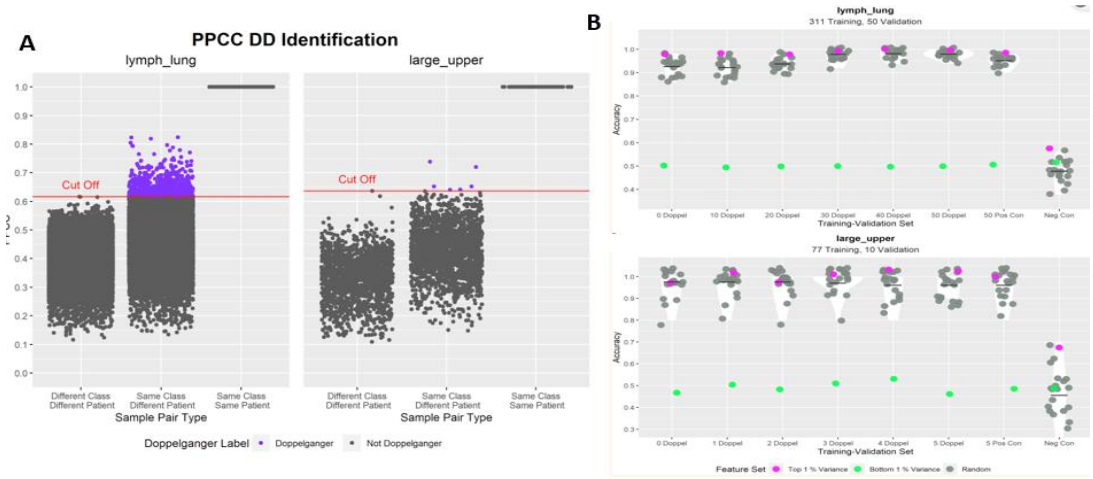

Figure 2(a).PPCC on lymph_lung and large_upper datasets. (b).Testing confounding effects of identified doppelganger data. (Wang et al. 2022)

4. Discussion

4.1 Ubiquity and Inevitability of data doppelgangers

Through the past researches on different proteome and genome database, we can conclude that doppelgangers effect exists widely among the genomic sequence database. The research on the doppelgangers effect of image-type medical data is still insufficient. Considering the deep learning network based on convolution algorithm is more widely used in image analysis, similar PPCC methods cannot be simply reproduced in medical image data. However, in the field of face recognition, similar studies have reported the doppelgangers effect of recognizing people with similar facial features as the same person. This indicates that the doppelgangers effect may exist in the image data, and that the doppelgangers effect may not be limited to medical data.

Considering that Wang et al. (2021) defined the doppelgangers data as samples with PPCC value exceeding any data pair of different classes, the universal existence of the doppelgangers effect is intuitive. If samples of different classes are considered to come from different distributions that can be classified, a higher relative coefficient from the same distribution can be reasonable. Statistically, the probability of high correlation of samples occurring from different distributions is often lower. Therefore, there is a high probability of the occurrence of doppelgangers data that meets the definition, and machine learning model can be expected to give the same prediction for similar inputs which may explain the inevitability of the doppelgangers effect.

4.2 How to evaluate the doppelgangers effect

If it is considered statistically necessary of samples with high similarity occurring in the same distribution, more complex methods should be considered when identifying doppelgangers data. For example, a representative sample in the distribution may

have a high similarity to all other samples, and marking it as doppelgangers data and excluding the representative features that may impair the model's ability to learn the features. One possible improvement is to compare the correlation coefficients of the same sample with all other samples in the data set. If a sample has a high correlation with only a few samples but a poor correlation with most of the other samples, it can be more confident to identify it as doppelgangers data. When evaluating the impact of the doppelgangers effect, the result of the validation set at the case of data leakage can be used as an important evaluation index. If the result for a high proportion of doppelgangers data is similar to that of data leakage, it can be considered that the double effect leads to overestimation of model performance.

4.3 How to deal with the doppelgangers effect

If we consider that doppelgangers effect is normally to occur in random samples, we have to rethink whether it should be treated as an abnormal phenomenon. Manually excluding the doppelgangers data may change the data distribution between the training and validation set, and impact the fitting effect of the model. As Wang et al. (2021) proved that dividing all the doppelgangers data into training sets or verification sets will lead to the decline of model learning ability. Therefore, as suggested by Wang et al., establishing a hierarchical verification set with different proportions of doppelgangers data may be a more prudent way to evaluate the doppelgangers effect. In addition, the general datasets extract features through sequence structure, which is difficult to identify those sequences with different structures but similar functions (Wang et al., 2021). This requires interdisciplinary prior knowledge and further optimization of the feature extraction algorithm, or the implementation of a large amount of data and the extraction of features through unsupervised learning encoder.

5. Reference

1. Garg, A., & Mago, V. (2021). Role of machine learning in medical research: A survey. Computer Science Review, 40, 100370.
2. Richard Webster, B., Hu, B., Fieldhouse, K., & Hoogs, A. (2022). Doppelganger Saliency: Towards More Ethical Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2847-2857).
3. Wang, L. R., Wong, L., & Goh, W. W. B. (2021). How doppelgänger effects in biomedical data confound machine learning. Drug Discovery Today.
4. Wang, L. R., Choy, X. Y., & Goh, W. W. B. (2022). Doppelgänger spotting in biomedical gene expression data. Iscience, 25(8), 104788.
5. Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016). The Doppelgänger effect: Hidden duplicates in databases of transcriptome profiles. JNCI: Journal of the National Cancer Institute, 108(11).