Doppelganger Effects in Machine Learning

1. Introduction

In recent years, machine learning models have been increasingly used to analyze biomedical data and diagnose diseases [1]. The important features of a specific dataset can be learned by these models to help the researchers make predictions about other data. In this case, model selection plays a significant role. Better performance on the validation dataset demonstrates stronger robustness of models. However, due to the effect of data doppelganger, the ultimate result of model selection may be deceptive and misleading. This report will generally explore if the doppelganger effects are unique to biomedical data. Meanwhile, to figure out some ways to avoid and eliminate the negative effects of data doppelganger in the future.

2. Doppelganger effects on biomedical data

The data doppelganger refers to the training and validation sets share high similarities. According to Wang's (2021) research, if a classifier falsely performs well on validation set due to the effect of data doppelgangers, there is an observed doppelganger effect [2]. In other words, the classifier may remember the important features extracted from training set, which also exist in the validation set. This means that the training set and validation set share the similar features. For example, in a model training process, there are 50 training samples, 20 validation samples and 30 testing samples. Through the data exploration, there are 18 training samples that share high similarities with 18 validation samples. In this case, when training the classifier with in-hand training data, and evaluating the model performance with validation data, the validation accuracy rate will be constantly greater than or equal to 18/20. The result cannot prove the current classifier is the optimal choice. It only illustrates that the classifier remembers the 18 similar features in the training process. However, the accuracy on the testing set is lower than the training accuracy because the testing data is totally new to this classifier. The classifier does not learn sufficient and accurate data features in the training process. Doppelganger effects have a negative impact on model selection. The selected classifier with high accuracy on the validation set has limited ability to generalize well on the testing set.

Biomedical data can be obtained in a variety of forms from a wide range of sources [3]. For instance, the signals, images and laboratory data from EEG, MRI scans and blood respectively. As Figure 1 shows, three types of immature granulocytes. The machine cannot identify the differences in terms of their size and complexities. When building the training dataset with such images, the machine cannot differentiate the data, even with human labeling. During the evaluation section, the automation will still treat them as the same label, which needs to be artificially judged to distinguish different leukemias or myelomas (UTS, subjects PowerPoint).

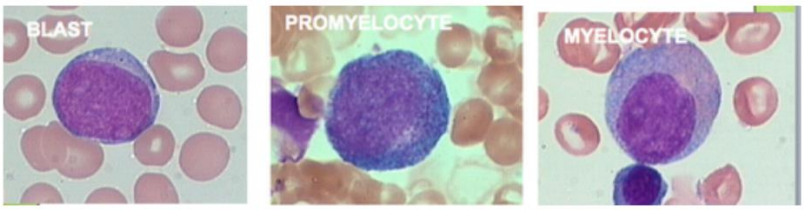

Figure 1. Screenshot of Blood smear (UTS, subjects PowerPoint)

In addition, the Electroencephalogram (EEG) signals are also typical biomedical data, which is challenging to classify and extract the features due to their characteristic of being densely non-linear and similar to each other (Figure 2)[4]. Hence, the classification of EEG signals has been an important research topic for nearly 30 years.

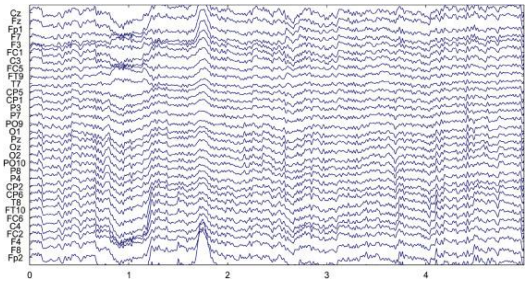

Figure 2. EEG data with 32 channels [5]

Besides, for face recognition system, high similarity faces will increase the probability of false matches. Depends on the research [6], biologically unrelated lookalikes have been reported to cause false matches either. As Figure 2 presents, each pair (column-wise) of the face looks very similar to each other.

Figure 3. Example of doppelgänger image pairs [6]

Data doppelganger effects may be a common problem for biomedical data. However, doppelganger effects are not unique to the biomedical data. Theoretically, given a dataset with many features, if some of these features bring overlapping information to each other and represent the same situation, this can be called data doppelganger. For the high-dimensional data, this is a common problem as well. To extend the face recognition system into the image identification field, there must be two images that share the high similarities, which leads to a doppelganger effect.

3. Eliminate doppelganger effects

In the classification problem, higher accuracy usually represents better performance. Doppelganger effects make the final result of machine learning models unreliable, higher accuracy may be generated by high-similarity training and validation set. Therefore, it is necessary to eliminate the doppelganger effects. To identify the presence of data doppelgangers between training and validation dataset is the first step. Wang (2021) pointed to a useful method to identify the data doppelganger. In the research, a pairwise Pearson's correlation coefficient (PPCC) was imported to measure the similarity within the selected batch. To observe the PPCC result, the paper firstly constructed negative, positive and valid cases with renal cell carcinoma. The PPCC data doppelganger depended on the PPCC distribution of valid cases against the negative and positive cases.

A simple way to avoid doppelganger effects is to place all PPCC data doppelgangers in the training set or validation set. The potential troubles are that the ultimate model may not generalize well if the data doppelgangers are all placed in the training set. In addition, after identifying the data doppelganger, removing all doppelgangers is another method to avoid doppelganger effects. This is not recommended to the small size dataset.

In addition, according to the advice [2], implementing data stratification will be a good method to avoid doppelganger effects. To stratify test data based on different similarities, rather than evaluating model performance on the whole test set. For example, after identifying the data doppelganger by PPCC, the test data can be divided into PPCC data doppelganger and non-PPCC data doppelgangers and to evaluate model performance on each set separately.

Additionally, some nonlinear measures can be applied to replace the PPCC. PPCC can be used to measure the linear relationship between two data. However, in machine learning, most data are complex, especially biological data, which is high-dimensional and complicated. The data features may also exist in different spatial domains. Thus, PPCC cannot fully describe the relationship between the two data because it can only measure the linear information between any two data. When we use machine learning or deep learning methods, our objective is to find the difference between the features of two data. In other words, the process of machine learning is to find the distinguishable feature points from the high-dimensional spatial domain. To describe the high dimensional features, the nonlinear information of the data should be considered.

There are some methods that can be used to describe the nonlinear characteristics of data. One of the common methods is Mutual Information (MI). This method uses entropy to describe the information relationship between two data. Compared to PPCC, MI can better express the nonlinear relationship between features. Another method is to calculate the Kullback-Leibler (KL) divergence or the Jensen-Shannon (JS) divergence of two features [7]. KL divergence is a measure that is used to describe the matching degree of two probability distributions. The large KL divergence refers to the greater difference between the two distributions. KL divergence can better describe the feature distribution between two data. JS

divergence is similar to KL divergence, which can be used to describe the similarity of two probability distributions.

A more direct method than KL divergence is to calculate the Wasserstein distance between two data [8]. If two distributions are far apart and do not overlap at all, the KL divergence value is meaningless, while the JS divergence value is a constant. This means that the gradient of this point is 0. Thus, Wasserstein distance measures the distance between two probability distributions, and it can provide a meaningful gradient. Therefore, these methods can describe the relationship between two data from different perspectives. We can calculate the linear and nonlinear relationship between two data and balance these relationship variables by weights. Eventually, we can obtain a variable to describe the matching degree between two features to identify the data doppelgängers.

4. Conclusion

Data is one of the most important parts of machine learning. Although lots of researchers have proposed advanced machine learning algorithms, if there are problems with training data, the model may not be accurate. In this paper, we present the impact of data doppelgängers, especially in biological data. Some methods are also proposed to avoid doppelgängers. In the future, how to solve doppelgängers effects is a potential research direction. If we can effectively avoid doppelgängers, the model performance of machine learning may be significantly improved.

5. Reference list

[1]. K. R. Foster, R. Koprowski and J. D. Skufca, "Machine learning, medical diagnosis, and biomedical engineering research - commentary," *Biomed Eng Online,* vol. 13, pp. 94, Jul 5, 2014.

[2]. L. R. Wang, L. Wong and W. W. B. Goh, "How doppelgänger effects in biomedical data confound machine learning," *Drug Discov Today,* vol. 27, no. 3, pp. 678-685, Mar, 2022.

[3]. S. Myneni and V. L. Patel, "Organization of Biomedical Data for Collaborative Scientific Research: A Research Information Management System," *Int J Inf Manage,* vol. 30, no. 3, pp. 256-264, Jun 1, 2010.

[4]. S. H. Ling, H. Makgawinata, F. H. Monsivais, A. Dos Santos Goncalves Lourenco, J. Lyu and R. Chai, "Classification of EEG Motor Imagery Tasks Using Convolution Neural Networks," *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2019, pp. 758-761, Jul, 2019.

[5]. R. Ghosh, N. Deb, K. Sengupta, A. Phukan, N. Choudhury, S. Kashyap, S. Phadikar, R. Saha, P. Das, N. Sinha and P. Dutta, "SAM 40: Dataset of 40 subject EEG recordings to monitor the induced-stress while performing Stroop color-word test, arithmetic task, and mirror image recognition task," *Data Brief,* vol. 40, pp. 107772, Feb, 2022.

[6]. C. Rathgeb, D. Fischer, P. Drozdowski and C. Busch, "Reliable detection of doppelgängers based on deep face representations," *IET Biometrics,* vol. 11, no. 3, pp. 215-224, 2022.

[7]. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems,* vol. 27, 2014.

[8]. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. Courville, "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017.