Q3 Report

1. Introduction

Machine learning has been more and more widely used in biomedical data in recent years to provide better solutions for pharmaceuticals, biomedical imaging, and health care etc... (Pierre, 2018) Sometimes, we can get good machine learning model performance, but due to doppelganger effects, the actual result is not necessarily good. The reason for this phenomenon is because the doppelganger effect seriously affects model evaluation. Doppelgänger effects occur when there are many data doppelgängers in a set of data. In other words, when there are many similar data in both training dataset and validation set, doppelganger may occur. Since the appearance of doppelganger effects requires the above characteristics, I think doppelganger effects appear not only in biomedical data, but also in other types of data such as Solar panel image data. In order to reduce the inflationary effect caused by doppelganger, it is very important to take some measures in the development of models. In addition to methods suggested by the author in the article like careful cross-checks using meta-data, perform data stratification and so on, I also proposed feasible methods such as manual intervention to increase the number of dissimilar data in validation dataset to reduce the proportion of functional doppelgängers and apply adversarial validation in both datasets.

2. Doppelganger effects are not unique to biomedical data

From the article, I learned that data doppelgängers in biomedical data (renal cell carcinoma proteomics data) often occur in different patients but the same class of tissue. With this discovery, I did an analogy analysis. So, I was thinking that if certain group of data comes from same class but different objects, then this group of data has a high probability of containing a lot of data doppelgängers. Solar panels data is a good example. In a solo panel dataset, there are 2,624 solo panel images from 2,624 solo panels of the same class, for each solo panel, there are 44 different solar modules (Zhang, 2021). According to Figure 2.1, we can see that the shade of colour represents the degree of damage to the solar panel. Through some pictures of the dataset, I found that some modules of the solar panel are generally damaged to a high degree, and some modules are generally damaged to a low degree. This is reasonable because some modules may have design flaws. According to this finding, when we want to use machine

learning algorithm for defect detection, if data in training and validation dataset are from the same modules but different solar panels, there is a high possibility that data doppelgängers occur. Therefore, I can easily deduce that when training this type of dataset (Some data in training dataset and validation dataset from same modules but different solar panels), the performance of the model will look relatively high, but in fact it has poor performance.

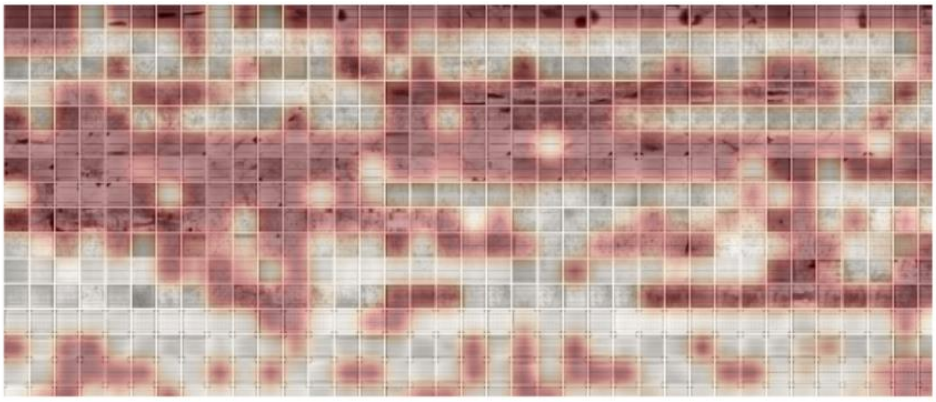

Figure2.1 Example of Solar panel image data

Not only image data like solar panel image may have doppelgänger effects, but non-image data may also have doppelgänger effects. If there is a strong correlation between different input data, the possibility of doppelgänger effects may also tend to be high. Suppose we need to predict a person's risk of stroke by parameters such as blood pressure, age, BMI, etc. According to Figure 2.2, blood pressure and age has positive correlation. So, we can deduce that if patients have similar age, blood pressure of patients is also more likely to be similar. This situation increases the similarity between the training dataset and the validation dataset data. According to the article, when the training dataset and the validation dataset are very similar, it is very likely that there will be a lot of data doppelgängers.

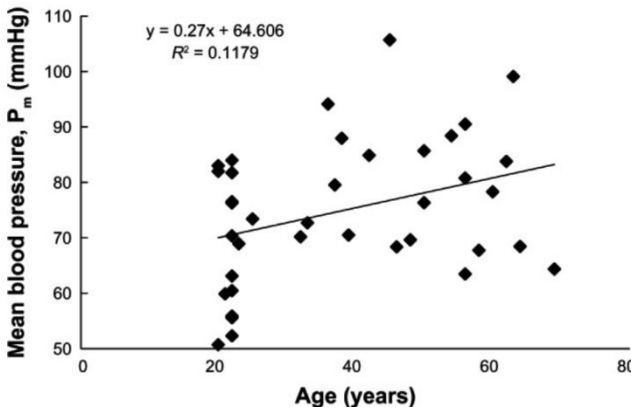

Figure2.2 The correlation between age and blood pressure

To conclude, doppelgänger effects are not unique to biomedical data, as long as the data in the training dataset and the validation dataset have a high similarity for some reason, doppelgänger effects may occur.

3. Methods to avoid doppelganger effects in the practice and development of machine learning models

In the article, the author provides several methods that may prevent doppelgänger effects, such as performing careful cross-checks using meta-data or perform data stratification Inspired by these recommendations of the author in the article, I propose some feasible methods to avoid doppelganger effects. The first approach is to increase the number of dissimilar data in validation data set to reduce the proportion of functional doppelgängers. under normal situation, the dataset is divided into training dataset and validation dataset in random ways. However, this segmentation method is not conducive to reducing doppelganger effects. Take the solar panel image data and age and blood pressure data as example, if we divide dataset in random ways, the data of the same solar panel module but different solar panel may appear in the training and validation datasets. We can manually add different modules data to the training dataset and validation dataset to reduce the similarity of data. Similarly, we can add some data into training and validation dataset that is not so correlated with age and blood pressure.

For the second approach, I propose to use Adversarial Validation to solve the Doppelganger effects problem. With Adversarial Validation, we can conversely think about the highly consistent distribution of training and validation data, and how to avoid this situation in the training model. Adversarial Validation is used to judge whether the data distribution of the training set is consistent with that of the validation set. In general, we use cross validation as the criterion for evaluating models, and then choose our final model. But in some data mining, the dataset is generally divided into training set and test set. For example, for many competitions dataset, due to data collection, sampling and distribution of training and online test datasets distribution may be inconsistent, so cross validation at this time can't accurately assess model on the test dataset.

The trick to mitigate this problem is to counter the Adversarial Validation. The variation of sample distribution is mainly reflected in the difference of data distribution between training set and test set. For example, men are increasingly appeared in the cosmetics or medical beauty market. Models based on data from the past are inapplicable to the present. For example, we

now recommend or predict the purchasing behaviour of Taobao users. Again, take the example of "predicting people's spending habits at the supermarket."(Figure3.1) Since the training set is mainly young people aged 18-25, and the test set is mainly old people over 70, we can distinguish the training set from test set by "age".

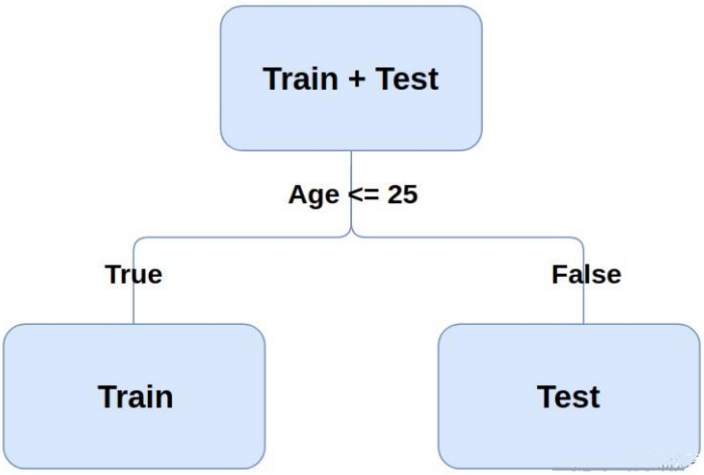

Figure 3.1 The classifier can easily distinguish training sets from test sets by age

Specific steps:

Define y: Is the sample train or test?

Combine Train and Test into a data set

Construct a model that fits the newly defined y

Observe the effect of the model: If the AUC of the model exceeds 0.7, it indicates that there is a large difference in the distribution of Train and Test

Instead of thinking consistency between training dataset and testing dataset, we consider inconsistency between training dataset and testing dataset. In the field of mutual finance, changes in market environment, adjustment of access strategy and quota strategy will lead to differences in the time dimension of training samples. Therefore, when training using large-scale samples or long-term samples, there must be bias between the training dataset and the testing dataset (or the current online scoring model). This phenomenon is reflected in the model index as the training set. The KS/AUC ratio on the verification set is higher, and decay naturally exists on the test set. In the face of this situation, we can not simply ascribe the reason to overfitting, but strive to improve the phenomenon of overfitting on the surface and need to consider the reasons for the changes in the above population.

In sum, Adversarial Validation method allow us to conversely think about the consistency between training dataset and testing dataset, but this method does not solve how to extract excessively similar data.

4. Conclusion

To conclude, doppelganger effects exist not only in biomedical datasets, but also in other types of datasets such as solar panel data. The appearance of doppelganger effects does not depend on the type of dataset, but on the characteristics of the dataset. When the training and validation datasets have a lot of similar data or training and validation dataset are consistent, doppelganger effects may occur. In addition, I propose some methods to avoid the doppelganger effects such as Adversarial Validation and increase the number of dissimilar data. However, these methods also have some problems. Both two methods can't remove similar data from training and validation dataset.

5. Reference

Baldi, P. (2018). Deep Learning in Biomedical Data Science. *Annual Review of Biomedical Data Science, 1*(1), 181-205. https://doi.org/10.1146/annurev-biodatasci-080917-013343

Brubaker, C., Jana, S., Ray, B., Khurshid, S., & Shmatikov, V. (2014). Using Frankencerts for Automated Adversarial Testing of Certificate Validation in SSL/TLS Implementations. *IEEE security & privacy, 2014*, 114-129. https://doi.org/10.1109/SP.2014.15

Qian, H., Wang, B., Ma, P., Peng, L., Gao, S., & Song, Y. (2021). Managing dataset shift by adversarial validation for credit scoring. arXiv preprint arXiv:2112.10078.

Qiu, L., Xiaojun, W., & Zhiyang, Y. (2019). A High-Efficiency Fully Convolutional Networks for Pixel-Wise Surface Defect Detection. *IEEE Access, PP*, 1-1. https://doi.org/10.1109/ACCESS.2019.2894420

Uangpairoj, P., & Shibata, M. (2013). Evaluation of vascular wall elasticity of human digital arteries using alternating current-signal photoplethysmography. *Vascular health and risk management, 9*, 283-295. https://doi.org/10.2147/VHRM.S43784

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