Data doppelganger Effect in Machine Learning

XIA Ri Xin,
x_rxin@163.com
Application Report for Msc Biomedical Data Science

1. Introduction

Data doppelgänger(DD) is defined as a high degree of similarity between the training and validation sets due to chance or other reasons. We say that there is a doppelgänger effect(DE) when a classifier incorrectly performs well due to the presence of DD. 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 data for Semantic Analysis. 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 using other correlation descriptors , considering more features or clustering methods.

2. DE are not unique to biomedical data

I learned that there are two key definitions – data doppelgängers (DDs) and functional doppelgängers (FDs), which can identify (or ‘‘spot’’) potential doppelgänger between and within data sets. DDs are sample pairs that exhibit very high mutual correlations or similarities. FDs are sample pairs that, when split across training and validation data, results in inflated ML performance. With this discovery, I did an analogy analysis. So, I was thinking that in semantic analysis, several semantics are output in English, Malay, Chinese, Tamil, etc. After feature extraction, the data distribution presents clusters formed by language rather than semantic clusters. For example in this case if the ML model is built using the training and validation sets of English language and tested using Malay language the results may be very bad, i.e. considering multiple languages the model is invalid.

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 DE in 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. In the following part, I will suggest some possible ways to avoid the dichotomy effect.

3.1 use other correlation descriptors

In the article, authors use the Pairwise Pearson's correlation coefficient (PPCC) to capture linear relations between sample pair of different data sets. An anomalously high PPCC value indicates that a pair of samples exists PPCC DD. Therefore I think other statistical correlation descriptors such as Pearson, Spearman, and Kendall type three correlation coefficients can also be tried as metrics for rating correlation.

3.1.1Pearson correlation coefficient (continuous variable)

$$\rho_{X,Y} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

$$\rho_{X,Y} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

3.2 Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}$$

Conclusion: Its value is not related to the specific values of the two variables of interest, but only to the size relationship between their values. di denotes the difference in position of the variables in pairs after sorting the two variables separately, and N denotes N samples, reducing the effect of outliers

3.3Kendall's tau-b rank correlation coefficient

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j)$$

Conclusion: It is a rank correlation coefficient. The rank correlation coefficient is 1 if the order is the same, and -1 if the order is the opposite.

the PPCC data doppelgänger identification method.

3.2 cluster methods

Clustering analysis is one of the most common analytical methods in machine mining. I think an attempt can be made to reduce DE by removing overly similar data

from such data through the idea of clustering. consider using the idea of greedy, starting with the longest input sequence as the first representative of the cluster, and then processing the remaining sequences from long to short, classifying each sequence as redundant or representative based on its similarity to the existing representatives. As inspired by data doppelgängers, model building and training can be performed using both data with excessive similarity removed and the original data, and then accuracy can be compared, to some extent, to reflect DEs.

3.3 Consider more features

In the scenario of protein function prediction, it is possible that an identical function may be performed by other proteins with different structures, which may lead to errors or limitations in the machine learning model. one of the possible reasons for the occurrence of DE is due to the wrong selection of necessary features for the sample; therefore the model cannot learn from proteins with different structures but similar functions.

Thus it is strongly advised that the feature selection of the model should take into account as many factors as possible. In this case, the primal reason is the limited information of structure feature, so that it is not distinguishable by using only structure feature. For example, factors such as the physicochemical properties of the protein, and the order of its sub-amino acids can be considered. In particular, concentrations should be on the theory and mechanism of why such different structure can perform similar function. Once we can find an essential and more discriminative feature other than structure, the problem can be solved.

4. Conclusion

In general, the doppelganger effect can be a challenging problem in the practice and development of machine learning models in health and medicine. However, it is possible to reduce the impact of doppelganger effect and improve the accuracy and reliability of machine learning models in this field by eliminating data that are too similar, using machine learning techniques for analysis, and developing models that take into account multiple characteristic factors.

Reference

Wang LR, Wong L, Goh WWB. (2021). How doppelg¨anger effects in biomedical data confound machine learning, Drug Discovery Today.

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

L.Waldron, M. Riester, M. Ramos, G. Parmigiani, M. Birrer. (2016). The Doppelg¨anger effect: hidden duplicated in databases of transcriptome profiles, J Natle Cancer Inst, 108

Natural Language Processing (NLP) Semantic analysis -- text classification, sentiment analysis, intention recognition: https://blog.csdn.net/javastart/article/details/117752296