doppelgänger effects

Abstract

Data doppelgängers occur when independently derived data are very similar to each other, affecting the reliability of some validation methods. This effect has a prevalence in biomedical data. The report will analyze the doppelgänger effects on biomedical data and explain how to avoid it in the practice and development of machine learning models.

Introduction

Machine learning is playing a more important role in biomedical areas, such as drug development. However, the performance of some ML models can be influenced by doppelgängers effect. The term data doppelgängers mean that training and validation sets are highly similar because of chance or otherwise. The ML models will perform well regardless of the quality of training because of the presence of data doppelgängers and thus make the model evaluation inaccurate. And this is the doppelgängers effect.

Data doppelgängers in biological data

Data doppelgängers have been observed in modern bioinformatics. For example, in the discovery of protein function prediction, proteins with similar sequences are inferred to be descended from the same ancestor protein and have a similar function to that ancestor. However, there are some exceptions such as twilight zone homologs. They have similar functions but have less similar sequences. There is abundance of data doppelgängers in biological data. And data doppelgängers are not unique to biomedical data. In can be found in many other areas.

Avoid data doppelgängers

It’s essential to identify the presence of data doppelgängers before validation. Thus we can avoid using these data to disturb the model's training. These are some ways to identify the data doppelgängers, like using ordination methods to make principal component analysis or embedding methods to see how samples are distributed in reduced-dimensional space or by dupChecker, comparing the MD5 fingerprints of their CEL files. However, they are not feasible in practice. One method, called Pearson’s correlation coefficient (PPCC), captures relations between sample pairs of different data sets and is proven possible. And it has been tested that PPCC data doppelgängers (based on pairwise correlations) act as functional doppelgängers, meaning that these data will confound ML outcomes. After finding these data, we can put the PPCC data doppelgängers all together in the training set and the doppelgänger effect will be eliminated in this way. If the dataset size is big enough, the data doppelgängers can be removed. Sometimes the size of the training set is fixed, and then it may be solved by ending up with spectacular winner-takes-all scenarios.

Conclusion

Data doppelgängers are common in many areas and not unique to biomedical data. It will cause many problems and we should try to avoid it. As for preventing data doppelgängers , because there is no good way to alleviate the data doppelgängers effect, it needs to identify the data doppelgängers in advance to prevent inflation performance.