Data Doppelganger

Summary

The main purpose of this report is to discuss the generation of doppelgangers, the impact of doppelgangers on biomedicine, and speculate on ways to remove the influence of doppelgangers in the field of machine learning. The doppelganger effect, which affects many current machine learning models and could hamper future biomedical advances, probably stems from the inescapable similarity in biology[1]. The report makes a number of recommendations about the effects of doppelgangers, based on previous researchers' published papers and the individual's undergraduate studies.

Introduction

Today, machine learning is a multidisciplinary discipline, involving probability theory, statistics, approximation theory, convex analysis, algorithm complexity theory and many other disciplines. The study focuses on how computers can simulate or implement human learning behaviors to acquire new knowledge or skills. Machine learning plays an important role in a variety of fields, including medicine, artificial intelligence, and business analytics[2]. However, it should not be ignored that machine learning is inevitably affected by the data, including data doppelgangers. This concept can be reflect in biomedical data, and it affects the performance evaluation part of machine learning[3], which is not fully explained due to its uncontrolled occurrence. But this concept is probably not just the preserve of biomedicine. Similar to it is quantum entanglement (a quantum phenomenon) in quantum physics, the unexplained synchronicity that affects the study of quantum.

The creation of doppelgangers

For doppelgangers, the root cause is the accidental and inevitable similarity of the

data, since the possibility of the data itself being similar exists and cannot be avoided at present. As a result, no matter what attention is paid, people or machines will always learn in the training set using data similar to the validation set. Biological data such as proteins or genes can reflect the existence of double identity[3], and when studying the function of protein or gene sequences, similarity is greatly considered. For biological image recognition, there are the same problems. Human faces, for example, inevitably have special similarities, which should be a phenomenon in nature[4].

The influence of data doppelgangers in machine learning

Doppelgangers make the actual performance of most trained models lower than the predicted performance, which decreases the applicability and accuracy of the model. In fact, the model may have good performance for a class of problems or a specific range of requirements[3], but the reality is complex and broad. For diseases, the accuracy of judgment directly determines the subsequent treatment effect. Doctors may misjudge because the condition of the contacted case is similar to other patients, but the critical evidence is different, and machines may also misclassify because of the high similarity of the received data. For the research of machine learning in the biomedical field, this is a difficult problem, because there is currently no magic pill to remove this doppelgangers effect[3], which may hinder the breakthrough progress in this field. Examples include facial recognition systems that misjudge faces with high similarity[4], a machine learning model has different performance when they predict different disease gene[5] , and so on.

Possible solutions to the doppelganger

1. Consider marking the data in the validation set from the point of data processing, and then weighting the data with high similarity between the validation set and the training set (the weighting method can be changed for reference according to the amount of data, for example, there are a lot of data with high similarity, and the

evaluation importance of all these data only equal to one irrelevant data). This data is less important for performance evaluation than other data with low similarity. This approach may solve performance inflation.

2. Consider from the data itself, since the data dimension reduction is not useful[3], can consider the data dimension increase. Similar data is nothing more than the current dimension, the values are basically the same. If the data similarity is changed, the difference of the data needs to be increased. At present, the data used to determine heart disease includes age, healthy lifestyle, etc[6]. If we add the time spent looking at mobile phone every day, whether we have good sleep, etc. (we cannot determine whether it is related to heart disease), then, due to the large sample size, the original similarity may be greatly reduced after adding the new dimension data.

3.To solve the problem that doppelgangers may make it difficult to classify dissimilar data, we can try to establish a negative feedback loop, and the correct establishment of this loop will help enhance the robustness and accuracy of machine learning model. The examples that are difficult to classify or misclassified in the validation set will be evaluated and added to the training set for retraining. At the same time, high PPCC data in the original training set will be replaced. This method may balance the proportion of different similarity of data. May enhance the mutation adaptability of the new data later.

Conclusion

All in all, doppelgangers are now inevitable[3], and in large part biomedical. Doppelgangers have certain impacts on health, medicine and other biological studies, and most of them are negative. Therefore, eliminating their effects is also an important research direction. However, there is no applicable treatment method at present. It can try to attach decision weight to similar data to change its influence on performance evaluation. You can try to increase the data dimension so that otherwise similar data may become less similar; It can be attempted to influence training set data in turn through validation set, remove similar data and add new available data after

evaluation.

Reference

[1] 汤可宗, 张彤, and 罗立民. "基于个体相似性评价策略的改进遗传算法."计算机应用与软件 33.3 (2016): 236-239.
[2] 张润,and 王永滨."机器学习及其算法和发展研究."中国传媒大学学报(自然科学版)23.02(2016):10-18+24. doi:10.16196/j.cnki.issn.1673-4793.2016.02.002.
[3] Wang, Li Rong, Limsoon Wong, and Wilson Wen Bin Goh. "How doppelgänger effects in biomedical data confound machine learning."Drug Discovery Today(2021).
[4] Rathgeb, Christian, et al. "Impact of doppelgängers on face recognition: database and evaluation."2021 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2021.
[5] Le, Duc-Hau. "Machine learning-based approaches for disease gene prediction."Briefings in functional genomics19.5-6 (2020): 350-363.
[6] Learning, Machine. "Heart disease diagnosis and prediction using machine learning and data mining techniques: a review." Advances in Computational Sciences and Technology 10.7 (2017): 2137-2159.