Doppelgänger Effects in Biomedical Data: Introduction and Thinking

Yijiang Liu

1 Introduction

Machine learning has been widely used in the field of biomedical research as it is far more efficient than traditional methods and usually has no ethical issues. And the quality of machine learning models is largely affected by the biomedical data used, so it is crucial to pay attention to choosing and checking the data before building machine learning models.

Researchers have found that a widespread phenomenon called doppelgänger effect is frequently observed in biomedical data[1], which has a negative effect on models. In detail, models will perform well whatever training methods have been adopted due to the existence of data doppelgänger in biomedical data and this is certainly misleading and will make the models untrustworthy, so it's very important to make sure that these similar data are identified before training. Data doppelgänger cannot guarantee a doppelgänger effect and more research is required if we want to understand the nature of data doppelgänger and try to avoid them as most methods proposed to solve the problem are still not robust enough.

Doppelgänger effect is not unique in biomedical data, and this can be widely observed in many fields. I'd like to talk about some examples about doppelgänger effects and methods that might be able to be used to avoid them by my knowledge and what I've learnt from the article.

2 Examples of doppelgänger effects

2.1 Doppelgänger effects in chemistry

Chemistry is a subject largely depends on experiments, but the low-efficient and great danger of certain operate have made more and more researchers interested in solving problems with the assistance of computers, so Cheminformatics today is a very popular track in chemistry and an important task is to predict the function of unknown chemical molecules according to known molecules.

It's widely agreed in chemistry that structure determines function, so researchers have built some models and basic data of molecules like bond length, bond energy and bond angle have been taken into consideration. For a new molecule, researchers can calculate its data and use the model to predict its function.

I have done similar work as a final assignment about two years ago. I chose some molecules which are mentioned in a medicine magazine and then got their basic data with computer software and used the data to build a model. It is not a surprise that the model can explain why those molecules are effective and even be used to find a new molecule that has potential function. But I found that this will only work when the structures of molecules are quite similar, if a molecule that has very unique structure is used in the model, the result will not be very good.

I think this is a good example of doppelgänger effect, the model can work very well partly because the adopted chemical molecules are very similar, only differ in some functional groups, so the data also don't have many differences and the model is not trustworthy in real situation.

2.2 Doppelgänger effects in statistics

Another interesting example I know about Doppelgänger effects is an experiment I heard in the university. The researchers wanted to know the attitudes of the students towards reading so they came to the library to distribute anonymous questionnaire every semester. It's said that most students in my university love reading according to the research.

I believe the result is not very convincing for some reasons. Firstly, the sample is too small and students in the lab and dorm should also be considered. Secondly, students in the library have potential similarities: they like reading books or study hard. What's more, although research have been conducted in different semesters, there can be same students in different samples as the questionnaire is anonymous, which might cause doppelgänger effects.

2.3 Doppelgänger effects in gene sequencing

In the research towards gene sequencing, doppelgänger effects can occur. Researchers have found that doppelgänger effects exist in RNA-Seq and microarray gene expression data[2].

3 Possible methods to avoid doppelgänger effects

3.1 Attach code to the sample

Inspired by the method dupChecker, I think we can attach a special code to each item when we are creating a database, the code should be logical and related to the characteristics of the items so when they are used to build a model, researchers can compare the codes of different samples and know whether they are similar according to the degree of match of the code.

3.2 Check the sample before use

Also, we need to pay more attention to the sample we choose before building models.

If we choose samples randomly there might be many similar or even repeated data and the doppelgänger effects can have a huge influence on the final result, so it's crucial to make sure that the data in the sample in a normal range and the sample itself is safe.

3.3 Promote the algorithm

Perhaps another method can be used to avoid doppelgänger effects is to promote the algorithm used in machine learning as doppelgänger data is too common to be avoided in the source of data. Maybe we can design a new algorithm that can ignore the doppelgänger data automatically during the machine learning so that we don't need to spend time on identifying them anymore.

4 Conclusions

Doppelgänger effects are very common in biomedical data and have a big influence on the result of machine learning models. It's noticeable that doppelgänger effects also exist in many other fields, including gene sequencing, chemistry and biostatistics. It's vital to pay attention to doppelgänger effects, otherwise the models will perform well whatever tragedies have been adopted in the progress of training. The best way to minimize the influence of doppelgänger effects is to identify them before model validation because it is very hard to resolve doppelgänger effects.

References

(1) Wang, L. R.; Wong, L.; Goh, W. W. B. How Doppelgänger Effects in Biomedical Data Confound Machine Learning. *Drug Discov. Today* **2022**, *27* (3), 678–685.

(2) Wang, L. R.; Choy, X. Y.; Goh, W. W. B. Doppelgänger Spotting in Biomedical Gene Expression Data. *iScience* **2022**, *25* (8), 104788.