Report on the paper - How doppelgängers in biomedical data confound machine learning

The use of machine learning has accelerated the development in the biomedical field, it helps researchers improve their working efficiency and discover new treatments for diseases. However, when researchers adopted the machine learning method in their studies, they sometimes couldn't build a good model because of data doppelgangers. Data doppelganger is used to describe the high similarity of the training and validation sets. This phenomenon that happened by chance will cause a doppelganger effect and confound machine learning outcomes - people will get a poor model with high accuracy on the validation set. To explore further on this problem, the researchers of this paper dug into the sources and consequences of data doppelgangers and recommend some possible ways to weaken the negative effects brought by data doppelgangers.

In the paper, the researchers cited several recent research to prove that doppelgangers are abundant in biological data. In my point of view, I agree that there is a greater possibility of being affected by doppelgangers in biomedical data. Firstly, considering the complexity of biochemical compounds and their various functions, it is difficult for us to label all the data concisely and make them easy to distinguish at the same time. On contrary, when we are doing other machine learning tasks like computer vision and natural language processing, we have a developed label system and could understand the result straightly. Secondly, there are still many mechanisms of macromolecular compounds such as Protein and RNA that remain unknown and we probably classify them based on their external functions even though there might be a huge difference hidden behind them. Thirdly, the exchange of chromosomes might create substances that are independently derived but very similar to each other, which is decisive to doppelganger effects. Thus, doppelganger effects will be more common in biomedical data. However, I don't think doppelganger effects are unique to biomedical. In other training tasks, if there are some duplications in the data sets, there is the possibility that leads to doppelganger effects.

After discovering the problem and proving its negative confounding effects, the researchers begin to look for ways to avoid data doppelgangers. In the beginning, they made several attempts to ameliorate data doppelgangers. They tried to split training and test data based on individual chromosomes, remove the PPCC data doppelgängers and trim the data by removing variables contributing strongly toward data doppelgängers effects. Unfortunately, non of these approaches worked perfectly and they all have obvious drawbacks. For the first approach, it is difficult to do practically because of the lack of prior knowledge and good-quality contextual/benchmarking data. The second approach, which does not work on small data sets with a high proportion of PPCC data doppelgängers. For the last approach, the researchers observed no change in the inflationary effects of the PPCC data doppelgängers after the removal of correlated variables because of the extreme complexity of the doppelganger effect.

Removing data doppelgangers from data is indeed complicated and elusive. The researchers gave us recommendations in the last part of the paper and pointed out ways for future studies. In my opinion, using multiple labeled data sets could be a way to avoid data doppelgangers. I got this inspiration while reading the identification part. The researchers attempt to detect data doppelgangers in a reduced-dimensional space. Though we can't distinguish the subtle differences in the current dimension, we could amplify the difference by adding a new dimension. Two similar data could possess a completely irrelevant label if we view them from another perspective, and this difference might become a great help to avoid data doppelgangers.

However, there are several difficulties in my approach. It is hard to find other dimensions which could help us to perfectly distinguish all the differences in similar data and adding an extra label will defiantly increase the training time significantly.