doppelgänger data and effect

Introduction

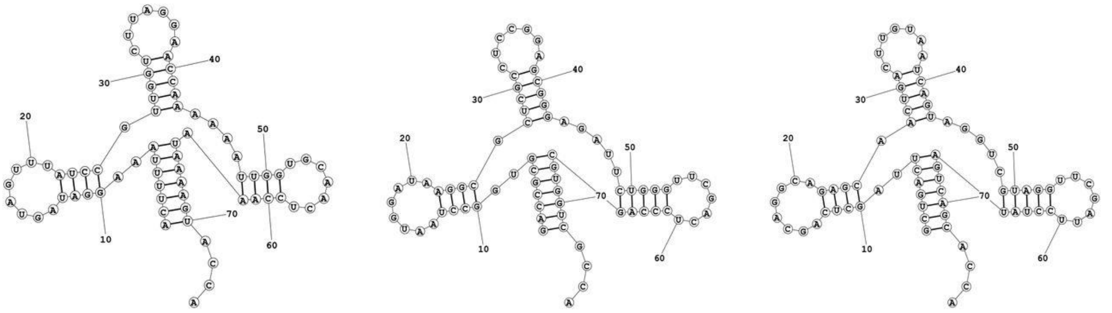
The paper describes what the doppelgänger effect is, and why it appears in biomedical data. This presentation will briefly describe what is covered in the paper. It will then present other data where the doppelgänger effect may occur, and then describe why doppelgänger data and the doppelgänger effect may exist in other fields as well. the final report will suggest possible solutions.

A brief description of the paper

The biomedical industry is using machine learning models to accelerate drug discovery and reduce the cost of testing drugs. doppelgänger effect is when a classifier falsely performs well because of the presence of highly similar training and validation sets. Some examples of the doppelgänger effect in biomedicine are also given in the paper. For example, quantitative structure-activity relationship models are trained classification and regression ML models for predicting the biological activity of molecules from their structural properties. Classifying similar molecules with similar activity into the training set and validation set confounds model validation, as poorly trained models may still perform well on these molecules. The paper mentions the use of paired Pearson correlation coefficients to identify doppelgänger data. However, so far there is no solution to the problems caused by the doppelgänger effect.

RNA secondary structure sequence prediction

In the past period, there are many researchers using machine learning to improve the prediction of new RNA secondary structures. However, for intra-family RNA secondary structure sequence prediction, the performance is not sufficient to prove its generality (Greener et al., 2022). The genes within the family are similar and a large portion of similar data exists for RNA. So it leads to the model training with similar data in the training and test sets, and the model performs well but has poor generalization ability.



(a) tRNA tdbR00000247 (b) tRNA tdbR00000372 (c) tRNA tdbR00000435

Figure1. Secondary structure of three tRNAs. Despite relatively low sequence identity (<60%), their secondary structures appear nearly identical. Many machine learning model benchmarks fail to separate these RNAs between the training and testing sets, causing significant overlap.

The doppelganger effect in the field of chemistry

According to the reasons for the formation of the doppelganger effect, doppelganger effect is not unique to biomedical data. The doppelganger effect also occurs when a large amount of similar data is available for other domains. For example, in the field of chemistry, the doppelganger effect also exists.

I) The doppelganger effect in the field of chemistry

The chemist Joseph Proust laid the foundation of modern chemistry, namely the law of constant composition. Chemical substances have a fixed composition. A chemical reaction has to take place with or in a substance to determine the chemical properties of the substance.

II) Examples of the doppelganger effect

In predicting the chemical properties of substances, substances with similar chemical elemental composition are used to infer that they have similar chemical properties. As cited in the paper, in protein function prediction, errors arise when proteins with similar sequences are inferred to be from the same ancestor. Therefore, when predicting the chemical properties of substances, this approach will not correctly predict substances that are not similar in chemical composition but are similar in chemical properties. For example, substances composed of isotopes have similar chemical properties despite their different compositions (K. Massila at al,2010).

How to eliminate the doppelganger effect speculation

This report conjectures that data pre-processing can be performed to transform similar data into different data to avoid doppelganger effect. For example, the prediction of protein function mentioned in the paper, the similar sequences in the data are hidden and only the different sequences are fed into the machine learning model to predict the protein. In the example of predicting molecular activity based on molecular structure one can consider hiding the similar parts of molecular structure and using different parts of molecular structure as data set and changing the machine learning algorithm to predict molecular activity.

Reference

Greener, J.G., Kandathil, S.M., Moffat, L., and Jones, D.T. (2022). A guide to machine learning for biologists. Nat. Rev. Mol. Cell Biol. 23, 40–55.

K. Massila, H. C. Soong and I. N. Haslinda, "Evaluation of Theoretical Isotope Generator as an alternative for isotope pattern calculator," *2010 International Symposium on Information Technology*, 2010, pp. 1051-1056,