The Doppelgänger Effect

Recently machine learning models have become increasingly important in the biomedical field and are currently being increasingly used in drug discovery to accelerate drug development. Machine learning improves the efficiency of drug discovery in many ways. By studying protein interaction networks, machine learning models can come up with better drug candidates faster, reducing the time spent on discovery and testing, and playing a crucial role in our human health. Cross-validation techniques are commonly used to evaluate these models. However, the reliability of such validation methods can be affected by the presence of data doppelgängers.

Whole-genome analysis of cancer specimens is commonplace, and investigators frequently share or re-use specimens in later studies. Duplicate expression profiles in public databases will impact re-analysis if left undetected, a so-called "doppelgänger" effect[1]. Several researchers have shown their ubiquity in biomedical data, demonstrated how doppelgangers are produced, and provided evidence that they affect experimental results.

According to information, there are two key definitions – data doppelgängers (DDs) and functional doppelgängers (FDs). DDs are sample pairs that exhibit very high mutual correlations or similarities. For example, we may use pairwise Pearson's correlation coefficient (PPCC) to identify DDs such that sample pairs with high PPCCs are also referred to as PPCC DDs. On the other hand, FDs are sample pairs that, when split across training and validation data, results in inflated ML performance, i.e., the ML will be accurate regardless of how it was trained (It can be assumed that such models have not truly "learnt")[2].

So far, there are few tools for doppelgänger identification or standard practices to manage their confounding implications, which has a great impact on our scientific research.

It seems to me that this phenomenon is not just in biomedical data, but in a wide range of other fields. Data doppelgängers have been observed inmodern bioinformatics. In one notable case, Cao and Fullwood performed adetailed evaluation of existing chromatin interaction prediction systems[3]. When they evaluated the system with a test set that was highly similar to the training set, they found that the system's performance was exaggerated. The presence of data doppelgängers was also observed by Goh and Wong, whereby certain validation data were guaranteed a good performance given a particular training data, even if the selected features were random.

In addition, data doppelgangers exist in the established field of bioinformatics: in

protein function prediction, proteins with similar sequences are inferred to come from the same ancestral protein and thus inherit the function of that ancestral protein. It tends to get better predictive performance, giving us a false impression of a highly accurate forecast.

In the study, the data doppelgängers effect may cause errors, so we need to identify whether doppelgängers exists between the training set and the verification set before validation, which may be the easiest way to avoid doppelgängers.

Earlier studies working on similar problems have also proposed measures for identifying data doppelgängers. One method, dupChecker, identifies duplicate samples by comparing the MD5 finger prints of their CEL files[4]. Identical MD5 fingerprints would suggest that samples are duplicates (essentially replicates and, therefore, indicative of leakage issues). Another measure, the pairwise Pearson's correlation coefficient (PPCC), captures relations between sample pairs of different data sets[5]. We used it to identify potential functional doppelgangers from the benchmark scenarios we built.

Recently, researchers have proposed a method for precisely matching duplicate cancer transcriptomes when nucleotide level sequence data is not available, even through different microarray techniques or through both microarray and RNA sequencing, which should be routine practice. They demonstrated the effectiveness of the approach in databases containing dozens of data sets and thousands of microarray profiles for ovarian, breast, bladder, and colorectal cancers, as well as matching microarray and RNA sequencing expression profiles from the Cancer Genome Atlas (TCGA). They found possible duplications in more than 50 percent of the studies, which came from different continents, used different techniques, were published years apart, and even within TCGA. This study may shed some light on how to solve this problem.

In the future direction of biomedicine, I believe machine learning and deep learning will play an increasing role. Now scholars have made a good connection between the two. Genome-wide association studies based on machine learning, for example, improve the scalability and efficiency of translating large imaging and textual data sets into phenotypes that can be used for genetic association studies, and could revolutionize the way we understand and manage genetic predispositions to disease. For example, using deep learning to help infer genetic variations from sequencer output that could be indicators of inherited diseases could greatly improve the diagnosis of diseases.

citations

[1] Waldron, Levi et al. “The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles.” Journal of the National Cancer Institute vol. 108,11 djw146. 5 Jul. 2016, doi:10.1093/jnci/djw146

[2] Wang, Li Rong et al. “Doppelgänger spotting in biomedical gene expression data.” iScience vol. 25,8 104788. 19 Jul. 2022, doi:10.1016/j.isci.2022.104788

[3] Cao, F., Fullwood, M.J. Inflated performance measures in enhancer–promoter interaction-prediction methods. Nat Genet 51, 1196–1198 (2019). https://doi.org/10.1038/s41588-019-0434-7

[4] Guo, T., Kouvonen, P., Koh, C. et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat Med 21, 407–413 (2015). https://doi.org/10.1038/nm.3807

[5] Levi Waldron, Markus Riester, Marcel Ramos, Giovanni Parmigiani, Michael Birrer, The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles, JNCI: Journal of the National Cancer Institute, Volume 108, Issue 11, November 2016, djw146, https://doi.org/10.1093/jnci/djw146