Report: How doppelgänger effects in biomedical data confound machine learning

One of the essential applications of the Machine learning (ML) model is drug research thanks to its efficiency, the test and discovery time, and cost reduced also it can be used for drug repurposing in other words to identify an existing drug to treat other diseases [1].

On the other hand, ML has a significant limitation, it will be affected by doppelgänger data. Thereby, identifying data doppelgängers between training and validation sets before validation. Although existing several methods, they are not generalized and robust. This study avoids the ordination methods and dupChecker which compares MD5 fingerprints in the CEL field to identify the duplicate sample. And choice the pairwise Pearson's correlation coefficient (PPCC) to analyze data.

Investigators used the renal cell carcinoma (RCC) proteomics data of Guo et al.9 taken from the NetProt software library20. To construct benchmark scenarios because it constructing clear-cut scenarios (i) negative cases, in which doppelgängers are not permissible by constructing samples pairs of different class labels; (ii) valid cases, in which doppelgängers are permissible by constructing sample pairs assigned to the same class label but from different samples. The resulting base on identified PPCC data doppelgängers based on the PPCC distribution of the valid scenario against the negative and positive scenarios has meaningful discrimination value. PPCC data doppelgängers represented in both training and validation sets inflated the ML performance.

In addition, this study used several methods to manage doppelgängers data, such as enforced colocation of doppelgängers in either training or validation sets and removing PPCC data doppelgängers or variables contributing strongly toward data doppelgängers effect but the results are unsatisfactory. Therefore, they suggest using meta-data to identify potential doppelgängers then prevent doppelgänger effects. And perform is data stratification is their second recommendation which avoids evaluating a model on whole test data but stratifies data into strata of different similarities. Finally, independent validation checks of huge data.

Doppelgängers effects not only present in biomedical; proteomics is commonly to present doppelgängers data in the protein prediction, it suggested that those proteins which have a similarity, have a similar function because have evolved from a common ancestor but this prediction by homology-based transfer [1] may be erroneous because different sequence could have similar function also protein from the same ancestor have a different function. For example, determination of the structure of type III effectors from *P. syringae* has yielded one novel fold (from AvrB), and another fold (from AvrPphF) that is sufficiently different from that of its closest relatives (ADPRT enzymes) to be no longer functionally identical [2]. In this case, elaborate a predict method based on a protein 3D structure, despite the sequence of proteins being similar but their 3D structure is different so this method will solve the problem.

Doppelgängers effects in gene sequence effect identification of human-associated bacterial species 16S rRNA sequence similarity is currently recommended to classify bacterially. As early as 1994, two strains were considered as belonging to distinct species if they shared 16S rRNA gene sequence similarities lower than 97% [3] and to discriminate two genera if this value was lower than 95%. In 2006, the cut-off value at the species level was re-evaluated at 98.7% (Stackebrandt & Ebers, 2006). Thereby, investigators analyzed to study the efficiency of this method. And the result is that phylum Chlorobi was the only phylum within which the 95 and 98.7% thresholds were respected for all studied taxa. Because 16 S rRNA could be variant and evolutionary. It demonstrated that classifying species based on 16S rRNA gene sequences may be insufficient. So, it may only be used as indicators, and not as a definite tool for the classification of bacterial strains. lead to misclassification and contribute to confusing situations.

Reference

1. L.R.Wang, L. Wong, W.W.B. Goh
How doppelgänger effects in biomedical data confound machine learning
Drug Discovery Today, (2021), ISSN 1359-6446, https://doi.org/10.1016/j.drudis.2021.10.017.

2. I. Friedberg
Automated protein function prediction—the genomic challenge
Briefings in Bioinformatics, Volume 7, Issue 3, (2006), pp. 225-242, https://doi.org/10.1093/bib/bbl004

3. D. Desveaux, A.U. Singer, J.L. Dangl
Type III effector proteins: doppelgangers of bacterial virulence
Current Opinion in Plant Biology, Volume 9, Issue 4, (2006), pp.376-382, ISSN 1369-5266, https://doi.org/10.1016/j.pbi.2006.05.005.

4. E. Stackebrandt, B. M. Goebel
Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology
Int J Syst Bacteriol 44 **(**1994 **)**, pp.846–849,
https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-44-4-846

5. E. Stackebrandt, J. Ebers.
Taxonomic parameters revisited: tarnished gold standards
Microbiol Today 33 **(** 2006 **)**, pp. 152–155,
https://ci.nii.ac.jp/naid/10020896069/