Paper Report

Title: How doppelgänger effects in biomedical data confound machine learning

Author: Li Rong Wang, Limsoon Wong, Wilson Wen Bin Goh

I. Purpose for the paper

When we use cross-validation techniques for ML model, data doppelgängers that occur when independently derived data are very similar will affect the reliability of such methods. However, although most methods of identify doppelgängers exist, most methods are not generalizable or robust enough. For investigating the nature of data doppelgängers, the authors show the prevalence of data doppelgängers in biological data, how it arise and provide proof of their confounding effects. The authors also put forward some methods to mitigate its effect.

II. Research methods

A. The authors use the below examples to show the prevalence of data doppelgängers in biological data :

   i. Cao and Fullwood: found that the performance of existing chromatin interaction prediction systems has been overstated and the test sets of these systems were highly similar to training sets;

   ii. Goh and Wong: also found the data doppelgängers;

   iii. The protein function prediction could not correctly predice functions for proteins with less similar sequences but similar functions;

   iv. If SAR paradox is the result of small variations in structure that substantially impact binding affinity, the poorly trained model would fail to identify the true biologicl activity.

B. The authors use the renal cell caecinoma (RCC) proteomics data of GUO et al and then be compared against positive cases to indentify data doppelgängers. They simulate these scenatios across the two batches of the RCC data set. (Fig. 1)

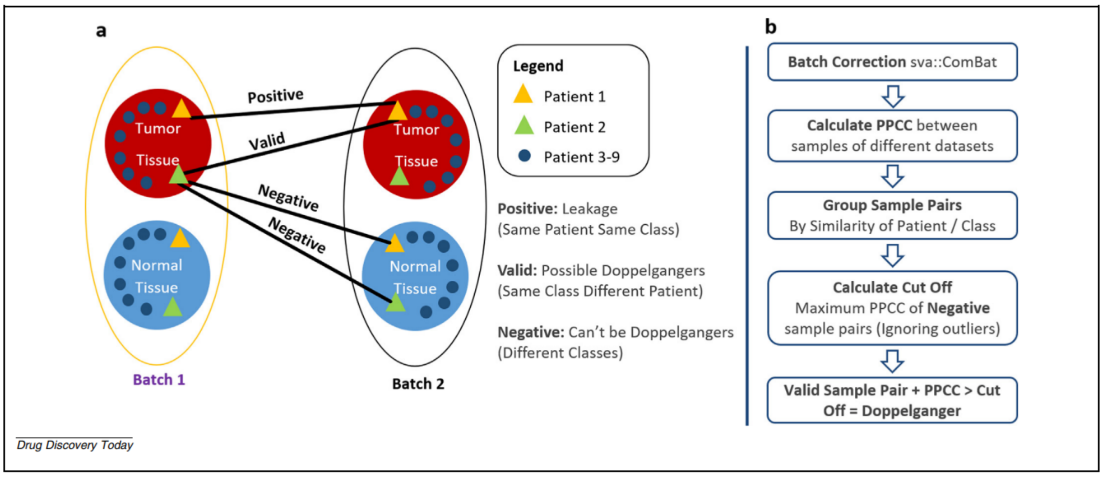
a
Positive
Valid
Negative
Negative
Tumor
Tissue
Normal
Tissue
Tumor
Tissue
Normal
Tissue
Batch 1
Batch 2
Legend
▲ Patient 1
▲ Patient 2
● Patient 3-9
Positive: Leakage
(Same Patient Same Class)
Valid: Possible Doppelgangers
(Same Class Different Patient)
Negative: Can't be Doppelgangers
(Different Classes)
b
Batch Correction sva::ComBat
Calculate PPCC between samples of different datasets
Group Sample Pairs By Similarity of Patient / Class
Calculate Cut Off Maximum PPCC of Negative sample pairs (Ignoring outliers)
Valid Sample Pair + PPCC > Cut Off = Doppelganger
Drug Discovery Today

FIGURE 1

Diagram illustrating the pairwise Pearson's correlation coefficient (PPCC) data doppelgänger identification method. (a) Naming convention for different types of sample pair based on the similarities of their patient and class. (b) Process of PPCC data doppelgänger identification. PPCC data doppelgängers are defined as valid sample pairs with PPCC values greater than all negative sample pairs.

Figure 1

C. The authors then explored data doppelgängers effects on validation accuracy across different randomly trained classifiers.

III. Creative ideas

A. The authors choose RCC data sets to identify data doppelgängers.

B. The authors find that ordination methods or embedding methods are unfeasible because data doppelgängers are not necessarily distinguishable in reduced-dimensional space.

C. DupChecker does not detect true data doppelgängers that are independently derived samples that are similar by chance.

D. The authors find the limitation of the original PPCC is it never conclusively made a link between PPCC data doppelgängers and their ability to confound ML tasks.

IV. Limitation of the methods

The authors attempted to alleviate doppelgänger effect with methods that would not lead to a significant reduction in sample size or require a high amount of contextual data, but fail.

V. Conclusion of the paper

A. PPCC has significant discrimination value, because in the secnarios, they observed a high proportion of PPCC data doppelgängers when replicates from the same sample or tissue is considered.

B. The presence of PPCC data doppelgängers in both training and validation data inflates ML performance; the more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance. (The result is in Fig. 2)

C. When all PPCC data doppelgängers are placed together in the training set, the doppelgänger effect is eliminated.

D. The authors put forword some methods to guard against doppelgänger effects:

   i. Perform careful cross-checks using meta-data as a guide;

   ii. Perform data stratification;

   iii. Perform extremely robust independent validation checks involving as many data sets as possible (divergent validation);

iv. Look for subsets of a validation set that are predicted correctly regardless of the ML method used and avoid to use them.

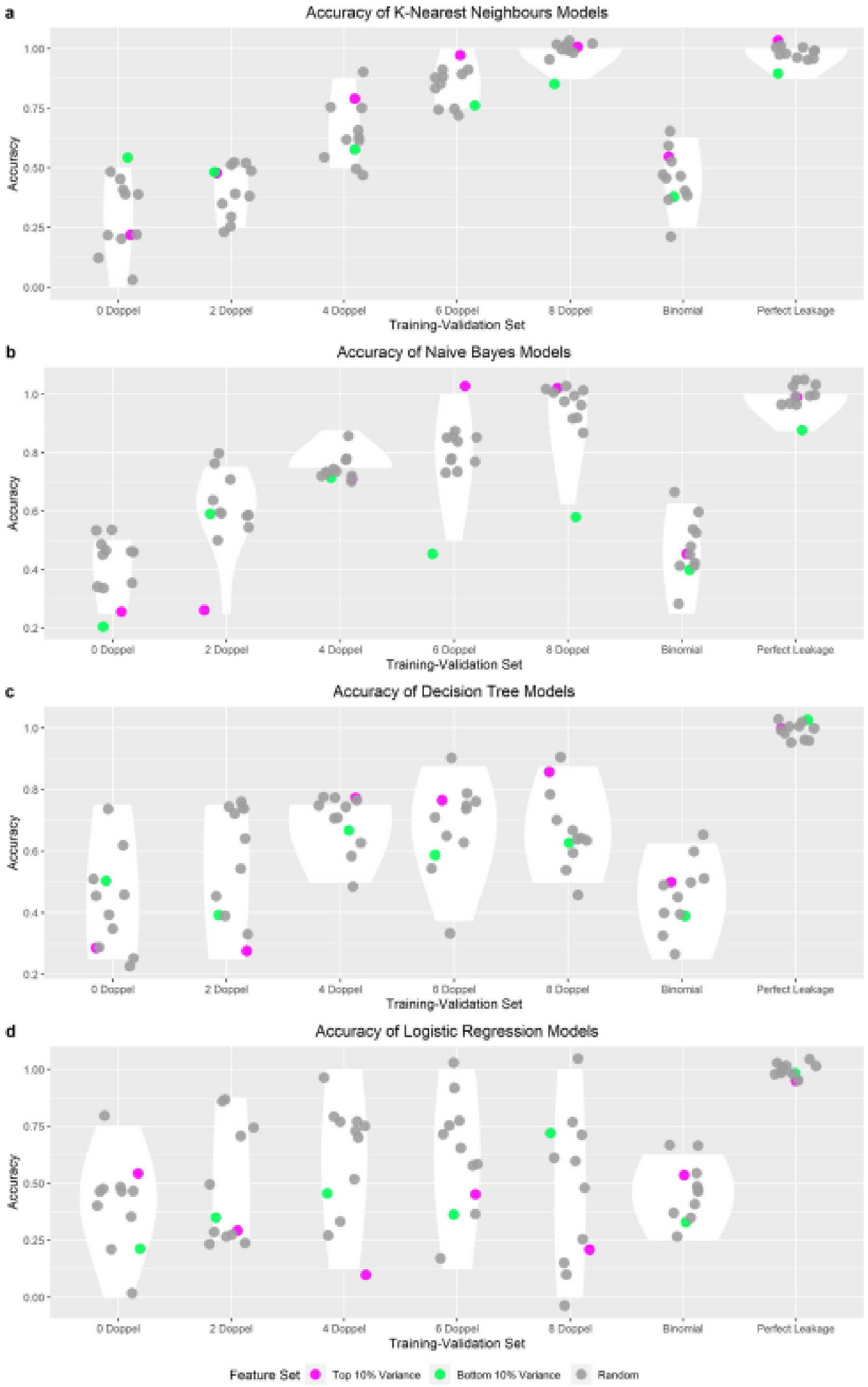

Figure 2

VI. Inspiration

I do not agree that doppelgänger effects are unique to biomedical data. Although most biomedical data is very similar to each other, I believe there are some other examples. For example, in the future, when the ML model is used to identify different similar dialects in the same province, the data sets may be very similar to each other, resulting in doppelgänger effects. According to the authors, there are some methods to avoid this effect in healthy and biomedical fields, such as checking original data sets carefully to divide them into different cases. Based on the fact that when all PPCC data doppelgängers are placed together in the training set, the doppelgänger effect is eliminated, we can identify potential doppelgänger effects using related information and then assort them into training sets or validation sets.