Report for the paper

‘How doppelgänger effects in biomedical data confound machine learning’

Introduction of Research

In this research, data doppelgangers are defined as ‘training and validation sets are highly similar because of chance or otherwise’(Wang, Wong, and Goh 2022). Data doppelgangers exist widely in biological data. In the context of machine learning models being increasingly applied to data in the biomedical field, data doppelganger may lead to models that perform well on training and validation data but actually have poor training quality, i.e., they do not perform well on validation data without data doppelgangers.

In addition, data doppelgangers may not necessarily have a doppelganger effect on the training of the machine learning model (i.e., confound the machine learning results). Data doppelgangers that have effect on the results of model training are termed functional doppelgangers. The main components of this research include illustrating the prevalence of functional doppelgangers under biomedical data, the impact of doppelganger data on ML, and methods to mitigate the doppelgangers effect.

Related Researches

Author/Researcher	Area of Data	Research
Cao and Fullwood	bioinformatics	The performance of existing chromatin interaction prediction systems has been overstated because of problems in assessment methodologies. These systems are evaluated on test sets that shared a high degree of similarity to training sets (Cao and Fullwood 2019).
Bin Goh and Wong	bioinformatics	Certain validation data were guaranteed a good performance given a particular training data, even if the selected features were random (Bin Goh and Wong 2019).
Wass and Sternberg	protein function prediction	naïve application of abductive reasoning is true in most cases (cases of data doppelgangers), but is unable to correctly predict twilight zone homologs(Wass and Sternberg 2008).
Friedberg	protein function prediction	naïve application of abductive reasoning is true in most cases (cases of data doppelgangers), but is unable to correctly predict enzymes that are dissimilar in sequence overall but with similar active site residues(Friedberg 2006).
Cherkasov et al.	drug discovery	Sorting similar molecules with similar activities into both training and validation sets confounds QSAR model validation because poorly trained models might still perform well on these molecules(Cherkasov et al. 2014).

Sheet 1: Related researches of data doppelgangers in biological data

Method	Note
ordination methods (e.g., principal component analysis)	unfeasible because data doppelgangers are not necessarily distinguishable in reduced-dimensional space
embedding methods (e.g., t-SNE), coupled with scatterplots	
dupChecker (identifies duplicate samples by comparing the MD5 finger-prints of their CEL files)(Sheng, Shyr, and Chen 2014)	does not detect true data doppelgangers that are independently derived samples that are similar by chance.
the pairwise Pearson's correlation coefficient (PPCC)	the prime limitation of the original PPCC paper was that it never conclusively made a link between PPCC data doppelgangers and functional doppelgangers. However, the basic design of PPCC as a quantitation measure is reasonable methodologically.

Sheet 2: Identification of data doppelgangers

Research process and conclusion

This research used renal cell carcinoma (RCC) proteomics data of Guo et al, taken from the NetProt software library, simulate scenarios across the two batches of the RCC data set. This research observed a high proportion of PPCC data doppelgangers (half of the samples are PPCC data doppelgangers with at least one other sample). It also checked PPCC distributions between same and different tissue pairs. PPCC values for same tissue pairs remain high overall, suggesting high correlations between samples, even if they come from different patients. These evaluations suggest that PPCC has meaningful discrimination value.

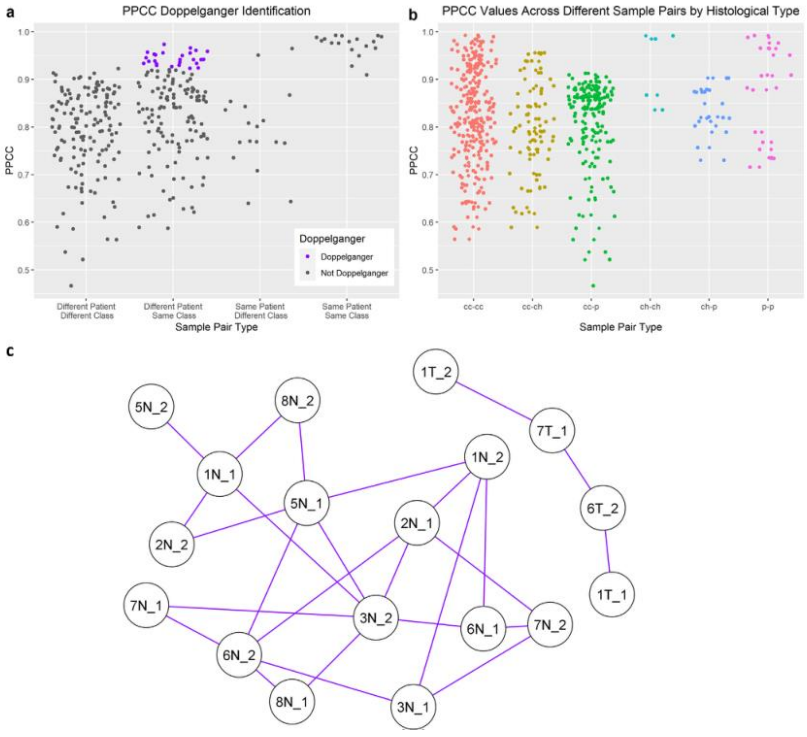

Figure 1: The data doppelgangers in PPCC

After identifying PPCC data doppelgangers in RCC, this research explored their effects on validation accuracy across different randomly trained classifiers. It was noted that the presence of PPCC data doppelgangers in both training and validation data inflates ML performance, even if the features are randomly selected (that means ML models would perform poorly on the data without doppelgangers). This finding is consistently reproducible on different sets of training and validation data and on different ML models. Moreover, the more doppelganger pairs represented in both training and validation sets, the more inflated the ML performance. This points toward a dosage-based relationship between the number of PPCC data doppelgangers and the magnitude of the doppelganger effect. This result confirms that PPCC data doppelgangers (based on pairwise correlations) act as functional doppelgangers (confounds ML outcomes), producing inflationary effects similar to data leakage.

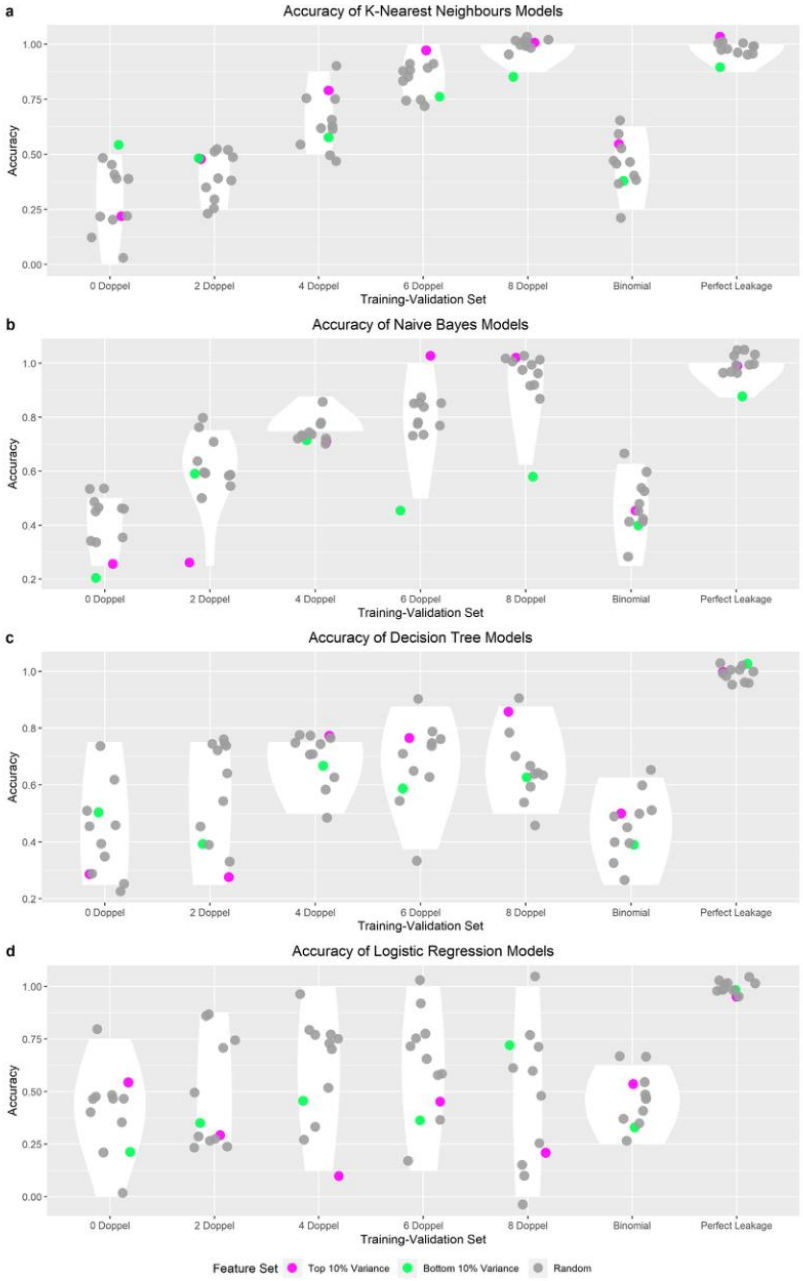

Figure 2: Performance of ML models (with data doppelgangers)

To eliminate the doppelganger effect, placing all doppelgangers in the training set is a possible way of avoiding the doppelganger effect. However, constraining the PPCC data doppelgangers to either the training or validation set are suboptimal solutions. In the former, when the size of training set is fixed (thus, each data doppelganger that gets included causes a less similar sample to be excluded from the training set), it leads to models that might not generalize well because the model lacks knowledge.

Author/Researcher	Method	Difficulty
Cao and Fullwood	called for more comprehensive and rigorous assessment strategies, based on the particular context of the data being analyzed(Cao and Fullwood 2019).	difficult to do practically because it predicates on the existence of prior knowledge and good quality contextual/benchmarking data.
Lakiotaki et al.; Ma et al.	doppelgangR(Lakiotaki et al. 2018; Ma et al. 2018)	does not work on small data sets with a high proportion of PPCC data doppelgangers, such as RCC, because the removal of PPCC data doppelgangers would reduce the data to an unusable size.

Sheet 3: The methods of removing data doppelgangers

Although removing data doppelgangers from data directly has proven elusive, this research still gave some recommendation to guard against doppelganger effects:
1: perform careful cross-checks using meta-data as a guide.
2: instead of evaluating model performance on whole test data, stratify data into strata of different similarities.
3: perform extremely robust independent validation checks involving as many datasets as possible.

My own thinking and discussion

From this research, I clearly understand that data doppelgangers exist widely in biological data and can seriously affect the training effect of machine learning models. The sample matching method used in this study is that data doppelgangers may exist in data of different patients from the same class. This demonstrates that data doppelgangers appear in samples with some of the same characteristics. Such samples may exist in fields other than bioscience, such as models for entity recognition. If an entity recognition model uses a training set from the chemical domain, its performance in recognizing non-chemical domain entities, such as geographic entities, must be worse. This is a rather extreme example, but to some extent we can consider that entities from the chemical domain have the same features, where data doppelgangers may exist. The way to avoid data doppelgangers as mentioned in the study, I think a more feasible approach at present is to perform cross-analysis of data before conducting data analysis to verify whether there is similarity in data from a metadata perspective. Performing robustness checks is also an effective method.

References:

Bin Goh, Wilson Wen, and Limsoon Wong. 2019. "Turning Straw into Gold: Building Robustness into Gene Signature Inference." *Drug Discovery Today* 24(1):31–36. doi: 10.1016/j.drudis.2018.08.002.

Cao, Fan, and Melissa J. Fullwood. 2019. "Inflated Performance Measures in Enhancer-Promoter Interaction-Prediction Methods." *Nature Genetics* 51(8):1196–98. doi: 10.1038/s41588-019-0434-7.

Cherkasov, Artem, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Victor E. Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. 2014. "QSAR Modeling: Where Have You Been? Where Are You Going To?" *Journal of Medicinal Chemistry* 57(12):4977–5010. doi: 10.1021/jm4004285.

Friedberg, Iddo. 2006. "Automated Protein Function Prediction—the Genomic Challenge." *Briefings in Bioinformatics* 7(3):225–42. doi: 10.1093/bib/bbl004.

Lakiotaki, Kleanthi, Nikolaos Vorniotakis, Michail Tsagris, Georgios Georgakopoulos, and Ioannis Tsamardinos. 2018. "BioDataome: A Collection of Uniformly Preprocessed and Automatically Annotated Datasets for Data-Driven Biology." *Database* 2018:bay011. doi: 10.1093/database/bay011.

Ma, Siyuan, Shuji Ogino, Princy Parsana, Reiko Nishihara, Zhirong Qian, Jeanne Shen, Kosuke Mima, Yohei Masugi, Yin Cao, Jonathan A. Nowak, Kaori Shima, Yujin Hoshida, Edward L. Giovannucci, Manish K. Gala, Andrew T. Chan, Charles S. Fuchs, Giovanni Parmigiani, Curtis Huttenhower, and Levi Waldron. 2018. "Continuity of Transcriptomes among Colorectal Cancer Subtypes Based on Meta-Analysis." *Genome Biology* 19(1):142. doi: 10.1186/s13059-018-1511-4.

Sheng, Q., Y. Shyr, and X. Chen. 2014. "DupChecker: A Bioconductor Package for Checking High-Throughput Genomic Data Redundancy in Meta-Analysis." *BMC Bioinformatics* 15(1). doi: 10.1186/1471-2105-15-323.

Wang, Li Rong, Limsoon Wong, and Wilson Wen Bin Goh. 2022. "How Doppelganger Effects in Biomedical Data Confound Machine Learning." *Drug Discovery Today* 27(3):678–85. doi: 10.1016/j.drudis.2021.10.017.

Wass, Mark N., and Michael J. E. Sternberg. 2008. "ConFunc—Functional Annotation in the Twilight Zone." *Bioinformatics* 24(6):798–806. doi: 10.1093/bioinformatics/btn037.