Reading Report of *How doppelgänger effects in biomedical data confound machine learning*

Machine learning described the process that gives computers the ability to learn or find out patterns without being explicitly programmed (Samuel, 1959). The application of machine learning is common in daily experience such as image recognition, speech recognition and traffic prediction. It has been increasingly used in computational biology such as bioinformatics, genomics analysis and biomedical engineering. For example, genomics machine learning can be trained to recognize the locations of transcription start sites (TSSs). The process of learning is to first develop a machine learning (ML) model then input a large collection of sequences that are either known to be TSSs or not TSSs—this collection of data is the training set. Then, a set of novel, undefined sequences are given to the model and will be predicted by the model whether they are TSSs or not (Libbrecht & Noble, 2015). This is the testing set. Finally, a validation set, which contains sequences independently derived from the sequences of the training set, will be given to the model to test the overall performance of this learning system (Libbrecht & Noble, 2015). If the model successfully predicts the label of the validation set, this learning system is considered successful. However, false validation on the performance of ML model could happen. When the data of training set and validation sets are highly similar by chance (even if they are independently derived), it will cause the model to intuitively perform well regardless of how poorly trained this model is. This is called the doppelgängers effect

(Wang et al., 2021). Data doppelgängers means that the data are highly similar like twins, and it would not always cause a doppelgängers effect. Functional doppelgängers are termed for the situation that data doppelgängers actually confound the ML model performance validation (i.e., a doppelgängers effect). In this paper, authors Wang et al. proposed a measure for identifying data doppelgängers as well as the confounding effects of the data doppelgängers identified by this measure.

The measure used in this paper is called the pairwise Pearson's correlation coefficient (PPCC), which captures the relations between sample pairs of different data sets (Waldron et al., 2016). A high PPCC value indicates that a pair of samples exhibit PPCC data doppelgängers (Wang et al., 2021). Authors constructed benchmark scenarios by using renal cell carcinoma (RCC) proteomics data of Guo et al.,2015. The RCC data was utilized into three sets: valid, positive, and negative. For RCC in negative sets in which the doppelgängers could not be observed since the samples are constructed with different class labels. Positive sets would always be doppelgängers permissible as the sample pairs are constructed by taking technical replicates from the same sample. Valid sets construct sample pairs with same class label but from different samples (Wang et al., 2021). Then the PPCC data doppelgängers was identified based on the PPCC distribution of the valid scenario against the positive and negative scenarios: PPCC data doppelgängers are defined as valid sample pairs with PPCC values greater than all negative sample pairs (Wang et al., 2021). Authors surprisingly observed a high proportion of PPCC data doppelgängers then expected,

suggesting that doppelgängers exist naturally as part of similarity spectrum between samples (Wang et al., 2021). However, the PPCC measure still has discrimination value when considering how many genes share common regulators in this scenario. To identify whether the determined PPCC data doppelgängers can cause functional doppelgängers effect, authors explored their effects on validation accuracy on different randomly trained ML models. They have noted that the presence of PPCC data doppelgängers in both training and validation data sets inflates the ML model performance in a dosage-based manner, the more PPCC doppelgängers pairs is present, the more inflated the ML performance. The result of this paper confirms that PPCC data doppelgängers will act as functional doppelgängers, causing the ML outcome inflated like data leakage (Wang et al., 2021).

Three recommendations were proposed by authors to reduce the doppelgängers effect on machine learning. First, perform careful cross-check using meta data as a guide. This allows plausible data doppelgängers are arising from same label class but different patients; Second, performing data stratification. Stratify data into PPCC data doppelgängers versus non-PPCC data doppelgängers; Third, performing extremely robust independent validation checks. Overall, identification of data doppelgängers is essential for machine learning system since the ML model performance assessment is only valid if validation data set are different and independent from training data.

Reference

Guo, T., Kouvonen, P., Koh, C. C., Gillet, L. C., Wolski, W. E., Röst, H. L., Rosenberger, G., Collins, B. C., Blum, L. C., Gillessen, S., Joerger, M., Jochum, W., & Aebersold, R. (2015). Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nature Medicine, 21*(4), 407–413. https://doi.org/10.1038/nm.3807

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and Genomics. *Nature Reviews Genetics, 16*(6), 321–332. https://doi.org/10.1038/nrg3920

Waldron, L., Riester, M., Ramos, M., Parmigiani, G., & Birrer, M. (2016). The doppelgänger effect: Hidden duplicates in databases of transcriptome profiles. *Journal of the National Cancer Institute, 108*(11). https://doi.org/10.1093/jnci/djw146

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

Samuel, A. L. (1959). Some studies in machine learning using the game of Checkers. *IBM Journal of Research and Development, 3*(3), 210–229. https://doi.org/10.1147/rd.33.0210