Report - How doppelganger effects in biomedical data confound machine learning

Zhiling Yan

December 7, 2021

1 Introduction

With the success of machine learning models implemented in medical field, scientists find the existence of data doppelganger phenomenon, even though the train and test sets are obtained independently. It is expected that the well-trained model performs better given that they are trained on informative structural properties, while poorly trained models fail in classification and regression tasks. However, because of the similarity between data-sets, doppelganger effect results in models performing well regardless of how they are trained [1].

Doppelganger effect has been observed in various bioinformatics cases, such as chromatin interaction prediction [2, 3], protein function prediction [4, 5] and drug discovery [6]. Data doppelgangers exert negative influence on conducting experiments, result analyses and model selection. Wang, Wong and Goh [7] applied pairwise Pearson's correlation coefficient (PPCC) to capture relations between sample pairs and to identify data doppelgangers. The result confirms that functional doppelgangers produce inflation in model performance, which is consistent with F. Cao et al [2]. It is also found that the overstated model with data doppelgangers can not show good performance in less related data-sets, such as the prediction of functions for proteins with less similar sequences but similar functions [8]. To solve the issue, related measures have been proposed to identify and eliminate data doppelgangers, such as principal component analysis (PCA) [9, 10], dupChecker [11], linear discriminant analysis (LDA), cross-check with meta-data [12], data stratification [13] and divergent validation [1]. However, the identification of data doppelgangers is still not accurate and how to remove the doppelgangers effects requires further study.

Inspired by Wang et al [7], I conducted experiments in Named Entity Recognition (NER), which is a basic task in natural language processing, and found results similar to doppelganger effects. The character-word lattice structure has been proved to be effective for Chinese named entity recognition (NER) by incorporating the word information [14]. By modify the injected word information, I can partly change the similarity between train and test sets. According to the obtained results, significant inflation occurred in all models, with the highest of 12.33% in Weibo data-set in FLAT model [15] and lowest of 4.41% in Ali e-commerce set in LEBERT model [16]. In addition, I found that the state-of-the-art model LEBERT overstated its performance because it injects extra word knowledge into the model, which means higher percentage of functional data doppelgangers in the input data.

2 Data doppelgangers in biomedical data

The discussion on data doppelgangers in biomedical data can mainly focus on three parts: the abundance of data doppelgangers in biological data, identification of data doppelgangers and how to eliminate data doppelgangers.

Abundance of data doppelgangers in biological data. As a fundamental mechanism in differential transcriptional regulation, enhancer-promoter regulation enables remote enhancers to contact with target promoters in cis to regulate gene expression [17]. And considerable methods are proposed to predict promoter-enhancer interactions (PEIs). However, Cao et al [2] found that train and test sets in these systems have a high degree of similarity, which results in overstatement of the performance of enhancer-promoter interaction-prediction methods. In addition, doppelgangers effect

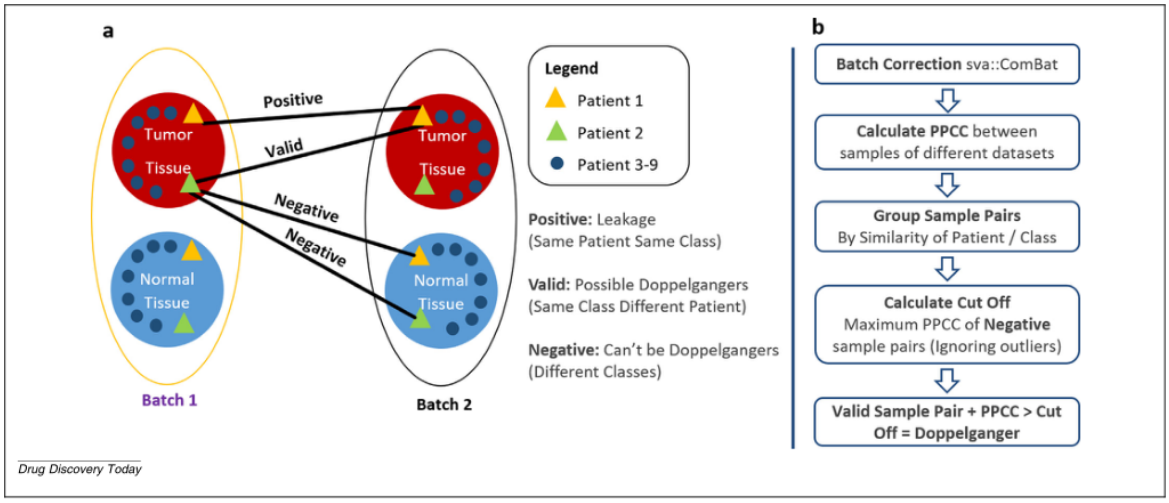

Figure 1: The structure of pairwise Pearson's correlation coefficient (PPCC) data doppelganger identification method.

occurs in automated function prediction (AFP). Though ontologies are currently the dominant accepted solution to this problem [8], they perform poorly in predicting functions for proteins with less similar sequences but similar functions, due to the existence of data doppelgangers in proteins with both similar sequences and similar functions. And data doppelgangers also occur in other cases, such as quantitative structure–activity relationship (QSAR) models, which is a computational modeling method for revealing relationships between structural properties of chemical compounds and biological activities [18].

Identification of data doppelgängers. Pearson's Correlation Coefficient is an important measure for capturing functional connectivity between different components [19]. For each pair of (x, y), Pearson's Correlation Coefficient is calculated using the following equation:

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

As a quantitation measure, pairwise Pearson's correlation coefficient (PPCC) is reasonable methodologically. So Wang et al applied this method to identify functional data doppelgangers in the renal cell carcinoma proteomics data [20]. The whole structure of method is shown in Figure 1. Firstly, they tagged sample pairs as negative, positive and valid samples according to the patient and class. Then, if valid sample pairs plus PPCC value were greater than cut off, maximum PPCC of all negative pairs, the data doppelgangers were identified. By analysing valid scenario against negative and positive scenarios, they successfully identified data doppelgangers and surprisingly found that 50% samples are similar to at least one other sample, which means a high percentage of data doppelgangers in the data-set. Inflationary effects were also observed in experiments with k-nearest neighbor models.

Preventing and eliminate data doppelgangers. To guard against doppelganger effects, Wang et al proposed three practical approaches: meta-data using , data stratification [13] and divergent validation [1]. The meta-data, as a high quality source of information [12], enables pathologists to efficiently identify and classify data doppelgangers and to prevent from data duplication. Besides, data stratification is considered as a valid method to prevent doppelgangers effects. For example, stratifying data according to types of cells enables less similarity among data-sets. And robust independent validation checks are also recommended.

3 Data doppelgangers in other data types

Named entity recognition (NER) is a classic task in natural language processing (NLP), which aims to locate and classify named entities in unstructured text into categories such as person names, organizations, locations, etc. Chinese NER is more challenging due to the lack of explicit word boundaries

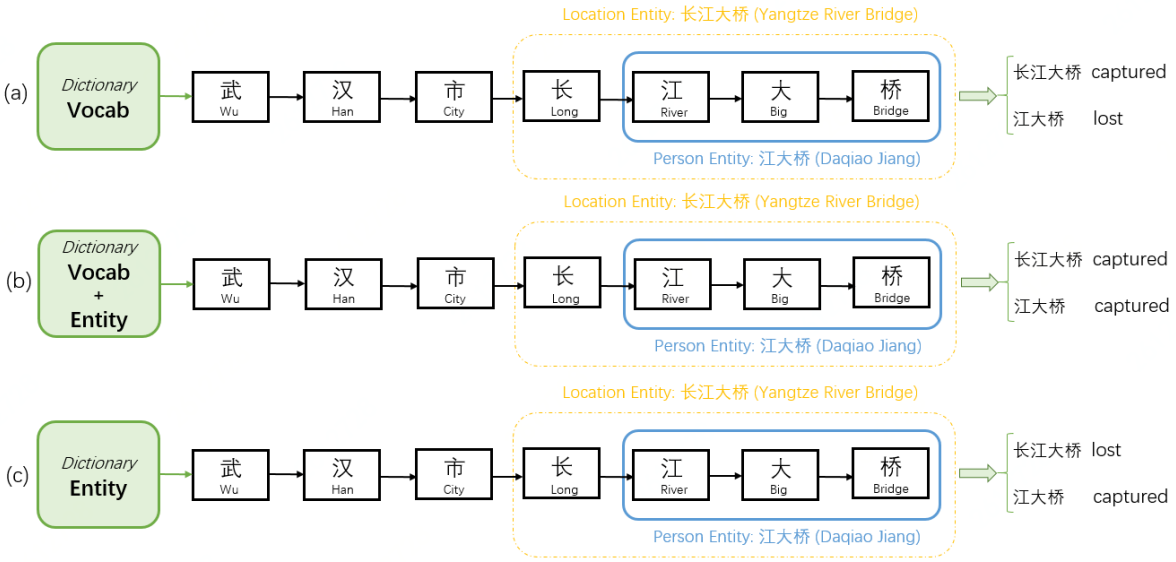

Figure 2: Diagram illustrating NER experiment.

Table 1: F1 score in Weibo.

Model	Vocab	Vocab+Entity	Entity
FLAT	60.67%	67.17%	73.00%
SimpleLexicon	56.16%	67.11%	67.64%
LEBERT	68.33%	71.51%	74.77%

Table 2: F1 score in Ali e-commerce.

Model	Vocab	Vocab+Entity	Entity
FLAT	74.11%	82.35%	86.31%
SimpleLexicon	71.23%	83.17%	85.39%
LEBERT	79.59%	81.31%	84.10%

in Chinese sentences [16]. It has been proved that lexicon-enhanced models for Chinese NER, which leverage word and word sequence information, perform better compared with character-based methods [14]. Therefore, the lexicon-enhanced models have been increasingly used in Chinese NER task these days [15, 21, 16, 22, 23, 24]. However, the output of my experiments reveals that the performance of LEBERT [16], the state-of-the-art model in Chinese NER, has been overstated due to the higher degree of similarity between training and testing sets compared with other baseline models. In addition, there is an increasing inflation in performance of all chosen models with more potential data doppelgangers according to my work, in both entity-level and dataset-level.

I found that data doppelgangers exist among different types of entities. I conducted this experiment using model BiLSTM+CRF [25] on Weibo set. As shown in Table 4, ORG means organization entities such as name of a museum, GPE means geopolitical entity such as China, Singapore. Because of the variety of ORG entities and high degree of similarity in GPE entities, F1 scores showed in these two types of entities are quite different from each other. This is much more similar to data leakage, since the same GPE entity occurs in both training and testing sets for higher probability compared with ORG entities. Using more types of entities to evaluate the performance of models is more robust and advisable . Most data-sets widely used in NER task contains at least three types of entity, such as four entity types in Weibo, eight entity types in Resume , three entity types in MSRA and four entity types in CoNLL.

Then, I worked on dataset-level, choosing the state-of-the-art model LEBERT [16] as well as two baselines FLAT [15] and SimpleLexicon [21]. The three models are all proposed to inject word information obtained from dictionary to enable data-sets to learn more lexicon knowledge. However, some entities in data-sets were not matched due to the limitation of the dictionary from papers called Vocab (as shown in Figure 2(a), gold entity "Daqiao Jiang" was lost). So, as shown in Figure 2(b), I added entities in data-sets into the dictionary and built Vocab+Entity dictionary. This time "Daqiao Jiang" as a word was captured in both training and testing sets, which means more potential similarity between the two sets. Although the word "Yangtze River Bridge" was captured from dictionary successfully, it was not a entity in the case, and resulted in extra noise in the data-sets. Therefore I conducted the third experiment shown in Figure 3(c). The dictionary including only entities enabled

Table 3: Entity coverage in dictionaries.

Model	Weibo	Ali e-commerce
FLAT/SimpleLexicon	58.35%	52.53%
LEBERT	86.97%	91.17%

Table 4: two entities of Weibo on BiLEST+CRF model

Entity	Number	F1 score
ORG	333	17.02%
GPE	361	51.76%

test set to share much higher similarity with training set with the noise being decreased to the bottom.

According to the work, I noted that the presence of potential data doppelgangers in both training and testing data inflated ML performance, even if the embedding vectors were randomly selected, consistent with results from Wang et al [7]. Moreover, more similarity represented in both training and testing sets, the more inflated the ML performance, with highest of 12.33% in FLAT and lowest of 6.44% in LEBERT from Vocab dictionary to Entity dictionary (shown in Table 1, 2). At last, the poor trained baseline model even performed better than the SOTA model (shown in Table 2: 86.31% of FLAT, 2.21% higher than LEBERT). In addition, I found that the state-of-the-art model LEBERT overstated its performance. LEBERT used bigger dictionary with higher degree of similarity between train and test sets compared with dictionary used in FLAT and SimpleLexicion. As shown in Table 3, dictionary applied in LEBERT contains 86.97% entities in Weibo and 91.17% entities in Ali e-commerce data-sets, while the dictionary used in both FLAT and SimpleLexicon covers half of entities. When I implemented the dictionary from FLAT SimpleLexicon to conduct experiment on LEBERT, I obtained result of 68.33% (shown in Table 1) in F1 score, 2.42% less than result in the paper.

4 Conclusion

Doppelgangers effects are observed in biomedical cases and result good performance even in poor trained models. I conducted a similar experiment in Chinese named entity recognition task by injecting extra lexicon knowledge and indirectly improving the degree of similarity between training and testing data. Inflationary effects occurred in both data-sets and entities. Performing robust validation checks may be effective to guard against Doppelgangers effects.

References

[1] S. Y. Ho, K. Phua, L. Wong, and W. W. B. Goh, "Extensions of the external validation for checking learned model interpretability and generalizability," *Patterns*, vol. 1, no. 8, p. 100129, 2020.

[2] F. Cao and M. J. Fullwood, "Inflated performance measures in enhancer–promoter interaction-prediction methods," *Nature genetics*, vol. 51, no. 8, pp. 1196–1198, 2019.

[3] W. W. B. Goh and L. Wong, "Turning straw into gold: building robustness into gene signature inference," *Drug discovery today*, vol. 24, no. 1, pp. 31–36, 2019.

[4] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, *et al.*, "A large-scale evaluation of computational protein function prediction," *Nature methods*, vol. 10, no. 3, pp. 221–227, 2013.

[5] M. N. Wass and M. J. Sternberg, "Confunc—functional annotation in the twilight zone," *Bioinformatics*, vol. 24, no. 6, pp. 798–806, 2008.

[6] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, and R. K. Tekade, "Artificial intelligence in drug discovery and development," *Drug Discovery Today*, vol. 26, no. 1, p. 80, 2021.

[7] L. R. Wang, L. Wong, and W. W. B. Goh, "How doppelgänger effects in biomedical data confound machine learning," *Drug discovery today*, 2021.

[8] I. Friedberg, "Automated protein function prediction—the genomic challenge," *Briefings in bioinformatics*, vol. 7, no. 3, pp. 225–242, 2006.

[9] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[10] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[11] Q. Sheng, Y. Shyr, and X. Chen, "Dupchecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis," *BMC bioinformatics*, vol. 15, no. 1, pp. 1–3, 2014.

[12] J. H. Caufield, Y. Zhou, A. O. Garlid, S. P. Setty, D. A. Liem, Q. Cao, J. M. Lee, S. Murali, S. Spendlove, W. Wang, *et al.*, "A reference set of curated biomedical data and metadata from clinical case reports," *Scientific data*, vol. 5, no. 1, pp. 1–18, 2018.

[13] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 145–158, Springer, 2011.

[14] Y. Zhang and J. Yang, "Chinese ner using lattice lstm," *arXiv preprint arXiv:1805.02023*, 2018.

[15] X. Li, H. Yan, X. Qiu, and X. Huang, "Flat: Chinese ner using flat-lattice transformer," *arXiv preprint arXiv:2004.11795*, 2020.

[16] W. Liu, X. Fu, Y. Zhang, and W. Xiao, "Lexicon enhanced chinese sequence labelling using bert adapter," *arXiv preprint arXiv:2105.07148*, 2021.

[17] A. Mora, G. K. Sandve, O. S. Gabrielsen, and R. Eskeland, "In the loop: promoter–enhancer interactions and bioinformatics," *Briefings in bioinformatics*, vol. 17, no. 6, pp. 980–995, 2016.

[18] R. Benigni, *Quantitative structure-activity relationship (QSAR) models of mutagens and carcinogens*. CRC press, 2003.

[19] T. Eslami, M. G. Awan, and F. Saeed, "Gpu-pcc: a gpu based technique to compute pairwise pearson's correlation coefficients for big fmri data," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 723–728, 2017.

[20] T. Guo, P. Kouvonen, C. C. Koh, L. C. Gillet, W. E. Wolski, H. L. Röst, G. Rosenberger, B. C. Collins, L. C. Blum, S. Gillessen, *et al.*, "Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps," *Nature medicine*, vol. 21, no. 4, pp. 407–413, 2015.

[21] R. Ma, M. Peng, Q. Zhang, and X. Huang, "Simplify the usage of lexicon in chinese ner," *arXiv preprint arXiv:1908.05969*, 2019.

[22] T. Gui, Y. Zou, Q. Zhang, M. Peng, J. Fu, Z. Wei, and X.-J. Huang, "A lexicon-based graph neural network for chinese ner," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1040–1050, 2019.

[23] R. Ding, P. Xie, X. Zhang, W. Lu, L. Li, and L. Si, "A neural multi-digraph model for chinese ner with gazetteers," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1462–1467, 2019.

[24] Y. Li, B. Yu, M. Xue, and T. Liu, "Enhancing pre-trained chinese character representation with word-aligned attention," *arXiv preprint arXiv:1911.02821*, 2019.

[25] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.