The Doppelganger Effects at Large

Potential doppelganger effect

Before presenting the final machine learning model, several candidate models are required to run on a test dataset to find out the best-performing model as the final model. However, the performance of a model may be exaggerated if the validation set and test set are highly correlated. In most cases, the prevalence of doppelganger effects lies in biomedical data due to unavoidable similarity among independently derived data.

Biomedical data may differ from other data types due to its high-dimensional meaning; similar problems can be discovered considering particular application scenarios, making the doppelganger effect unique to biomedical data sets.

There are striking similarities between linguistics and biology in some respects. Just like gene sequences from biology, our daily dialogues and texts can be processed and generate its 'gene sequence'— hash value. [1] Through some hashing algorithms, a long piece of data can be mapped to relatively short data, which is the hash value of big data. The hash value has the characteristic that it is one-to-one mapped to the big data, and once the big data changes, even a small change, its hash value changes accordingly. On the other hand, since the hash is the DNA of the big data, there will be no hash value exactly the same for any two pieces of data. Because of this characteristic, the Hash algorithm determines whether the two files are the same. Therefore, it is not surprising that doppelgangers may exist in the Neuro-Linguistic Programming (NLP) area.

In research on predicting the mental health of scholars using NLP [2], the researchers assumed that frustrated youngsters use certain curse words frequently on Twitter to identify frustrated individuals by checking the words they use. However, some curse words may be an indirect implication of anger or frustration; on the contrary, many young people like to express their excitement and cheers using words like 'high,' 'damn,' and 'rage,' which may confuse machine learning models as two complete contrary emotions are expressed in similar words. It is likely that the models presented in the paper may be exaggerated. And this problem is also mentioned by these researchers.

Apart from linguistics, machine learning can also be used in analyzing data in the entertainment industry. A logistic regression model designed to predict a movie's success before its release achieved an accuracy of 80% by mainly focusing on its music rating on different websites [3,4]. In most cases, most popular Bollywood movies have catchy background music, and music is an influential factor for the Indian audience. However, the sample size of selected movies is limited by adding this factor, which

may lead to the improved similarity between training and test data. Also, successful movie companies tend to invest more money in hiring skilled music producers, and therefore more critics and audiences may leave positive comments online.

High-dimensional characteristics of medical information

Diagnosis in modern medicine is based on a wide variety of information, including but not limited to medical history, genes, images (ultrasound, x-ray, ct, MRI, pet). Even if a lot of complete information is collected, analyzing data from a higher dimension is still incomplete since the human disease is a dynamic development process with time series characteristics. But pictures and other medical information are essentially a point in the time series that may not reflect the overall perspective. Because of individual differences, there may still be misdiagnosis as the whole data set is collected from different people.

For example, there are more than two thousand diseases in ophthalmology, and a patient has to do a lot of tests to determine which disease it is. At this stage, data-driven AI can only make one or two diseases diagnosis of binary classification (note that each disease also has different stages, such as glaucoma, and how many stages exist is a relatively open question). [5] Based on this situation, an AI is currently unable to predict diseases that it has not learned and can only give an answer to the most likely disease, which limits its application scenarios.

Avoiding and Ameliorating data doppelgangers

As mentioned above, elevating the dimension of original data might be an approach to reduce the doppelganger effect. Recently, a model can analyze multi-modal medical images to classify ophthalmic disease effectively by receiving both fundus images and OCT images in long-tailed distribution. And the researchers plan to input digital medical records with text information as a third modality. [6] In general, it is more likely that single-modal can make false predictions, whereas multi-modal models can focus on related information from similar pairs. Therefore, feature information can be extracted and analyzed more accurately in multi-modal models.

Inspired from the hash value in NLP, biomedical research has been accelerated by technological advances in biomedical text mining (BioTM), an approximate hashing for bioinformatics. [7,8,9] Using this technique, similarity among a large collection of DNA sequences can be quickly located, and the slightest dissimilarity can also be identified at the same time so that different weights can be assigned to features accordingly, which might be helpful to ameliorate doppelgangers. A hash subgraph

pairwise kernel-based approach for PPI extraction can outperform other systems on a very limited data set.[10]

Also, the hash value can be utilized in image processing and image retrieval. Shi et al. proposed a deep ranking hashing, converting medical images into binary forms. In this system, the distinction between different classes and in similar classes can be greatly emphasized.[11]

The future of machine learning in the biomedical area

What artificial intelligence (AI) needs to learn is understanding the link between extracted features and pathology, which is possible to push applications of AI in the medical area to a higher level. Modern medicine relies heavily on evidence-based research methods that are doomed to be significantly affected by human bias. Although it is possible to use statistics to reduce the bias caused in clinical trials, clinical trials themselves are limited by sample size and medical evidence. If AI can integrate the intuition of multiple doctors to maximize the rejection of bias and then promote pathological analysis in return, the development of health and medical science will be greatly accelerated.

Reference

[1] Hash Function[EB/OL]. [2021/11/27]. https://en.wikipedia.org/wiki/Hash_function.

[2] A. Chaurasia, S. V. Prajapati, P. A. Tiru, S. Kumar, R. Gupta and A. Chauhan, "Predicting Mental Health of Scholars Using Contextual Word Embedding," 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), 2021, pp. 923-930, doi: 10.1109/INDIACom51348.2021.00166.

[3] A. Kanitkar, "Bollywood Movie Success Prediction using Machine Learning Algorithms," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739693.

[4] S. R. Jaiswal, and D. Sharma, Predicting success of bollywood movies using machine mearning techniques, ACM Compute 2017, Nov, Bhopal India.

[5] Perdomo O., Andrearczyk V., Meriaudeau F., Müller H., González F.A. (2018) Glaucoma Diagnosis from Eye Fundus Images Based on Deep Morphometric Feature Estimation. In: Stoyanov D. et al. (eds) Computational Pathology and Ophthalmic Medical Image Analysis. OMIA 2018, COMPAY 2018. Lecture Notes in Computer Science, vol 11039. Springer, Cham. https://doi.org/10.1007/978-3-030-00949-6_38

[6] Z. Ou et al., "M2LC-Net: A multi-modal multi-disease long-tailed classification network for real clinical scenes," in China Communications, vol. 18, no. 9, pp. 210-220, Sept. 2021, doi: 10.23919/JCC.2021.09.016.

[7] G. Arbitman, S. T. Klein, P. Peterlongo and D. Shapira, "Approximate Hashing for Bioinformatics," 2021 Data Compression Conference (DCC), 2021, pp. 337-337, doi: 10.1109/DCC50243.2021.00072.

[8] A.M. Cohen and W.R. Hersh, “A Survey of Current Work inBiomedical Text Mining,” Briefings in Bioinformatics, vol. 6, no. 1,pp. 57-71, 2005.

[9] W. Hersh, A. Cohen, P. Roberts, and H.K. Rekapalli, “TREC 2006 Genomics Track Overview,” Proc. 15th Text Retrieval Conf. (TREC ’06), 2006.

[10] Y. Zhang, H. Lin, Z. Yang, J. Wang and Y. Li, "Hash Subgraph Pairwise Kernel for Protein-Protein Interaction Extraction," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 4, pp. 1190-1202, July-Aug. 2012, doi: 10.1109/TCBB.2012.50.

[11] X. Shi, M. Sapkota, F. Xing, F. Liu, L. Cui, and L. Yang, “Pairwise based deep ranking hashing for histopathology image classification and retrieval,” Pattern Recognition, vol. 81, pp. 14–22, 2018.