**Biological Doppelgangers**

**Introduction**

Data doppelgangers are independently derived data which are very similar to each other, which is commonly found in biomedical data (Wang *et al.*, 2022b). This phenomenon is not just restricted to medical field, they are used in many cases of machine learning application by big corporations and recommend the same products to the people with the same interests based on search (Zhang *et al.*, 2021). Similarly, in fraud detection systems, where the fraud and real transactions seem legitimate (Stojanović *et al.*, 2021). However, the ultimate difficulty of biomedical data manipulation lies in misuse of biomedical data and misunderstanding of machine learning modes in the case of treatment and diagnosis can harm the patient. The doppelganger effect is described as an inflated performance of a model due to repeated datapoints. This occurs through similarity of the datapoints and therefore overfitting of the model to the data. However, not all doppelgangers cause doppelganger effect and the ones that do are called functional doppelgangers (FG).

Different databases for specific disease (for example, cancer) may have samples from the same individuals taken independently, however, marked with different id. In such case, although the samples were taken independently, the first sample taken from the tumour would contain similar information as the second sample. As well as that, some samples can be processed in different ways using different software and use in different studies.

In biomedical research, doppelgangers pose a threat of inaccurate results. It may, perhaps, not cause any difference in analysis of the data, however, implementation of machine learning model on such data can cause inflation of the results. This is because the model can accidentally be overfit to the doppelganger samples. In addition, databases may share the expression data with other databases and the accumulated duplicates may be ignored which may affect future analysis.

When using repeats for machine learning it causes bias of the model overfitting for the repeated points. Here we are presenting an example of such overfitting:

1. Table

   Description automatically generatedResults of multiclass classification model (Random Forest) using a randomly generated dataset:
2. Table

   Description automatically generatedResults of the multiclass classification model (Random Forest) using a randomly generated dataset; however, the first fifty rows of that model are repeated ten times:

As we can clearly see from this, the results with the duplicated rows show clear inflation of the results, however, that would not have been so apparent have we not checked for duplicates. Therefore, pre-processing is a crucial step before any modelling problem.

**Detection of doppelgangers**

There is no universal method to remove doppelgangers, in fact, none of the existing methods are robust enough (Wang *et al.*, 2022b). Here we present some of the methods that could be found useful to avoid doppelgangers and their effects in machine learning.

1. Simply checking for duplicates

Checking and deleting duplicated rows can reduce overfitting of the model to the particular sample to improve the performance on unseen dataset.

1. Analyse correlations

Parsons corelation coefficient can help to remove the features with the highest correlation and give priority to the features that bring more value.

1. Reducing dimensionality with PCA

Principal component analysis or PCA can help to reduce dimensionality of the data and with this compressing the information to find patterns in the high dimensional data.

1. Specifying the stratification of the doppelgangers in different splits of Machine Learning and validation the results on an independent validation set

If the doppelgangers are present but you are confused about the possible tests that exists to assist the problem, the samples could be specifically stratify the samples, so they are evenly spread over train and tests splits. For better verification, external data of the same kind could be used as validation set for more robust testing of the model performance.

1. Specialised software for doppelgangers removal

Special software has been developed to remove the presence of the doppelgangers (Waldron *et al.*, 2016, Wang *et al.*, 2022a), which are a composition of above methods out into functions. It is a much easier, atomised methods which does the work for you.

After using these methods, we can use cross validation to evaluate our machine learning model. Cross-validation uses different splitting oof the data to train and test the model, making the model performance evaluation more objective. However, we can also try avoiding doppelgangers by developing robust checking systems to avoid their occurrence in the first place.

**Systems that could be developed to avoid the rise of doppelgangers**

Although there are a variety of methods to counteract the doppelganger effect, it is the primary concern to avoid the occurrence of doppelgangers before the data is manipulated. Usually, doppelgangers can occur when the databases are merged and samples from the same individuals, marked with different study identification number, are merged from two different studies. A more difficult case of that can also happen if the samples that were taken from the same were processed by different technologies and therefore when the samples are merged into one big database, they would not be a total duplicates, but still doppelgangers (Waldron *et al.*, 2016). This can be overlooked when the data is used for machine learning.

**Conclusion**

All in all, this mini report demonstrated the doppelganger effect on data, has discussed its effects on biomedical data and provided some methods that could be used to prevent inflated results of the model.

**Reference list**

Stojanović, B., Božić, J., Hofer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., Sundaram, M., Jordan, E.*, et al.* (2021) ‘Follow the Trail: Machine Learning for Fraud Detection in Fintech Applications’. *Sensors,* 21**,** 1594.

Waldron, L., Riester, M., Ramos, M., Parmigiani, G. and Birrer, M. (2016) ‘The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles’. *J Natl Cancer Inst,* 108.

Wang, L. R., Fan, X. and Goh, W. W. B. (2022a) ‘Protocol to identify functional doppelgängers and verify biomedical gene expression data using doppelgangerIdentifier’. *STAR Protoc,* 3**,** 101783.

Wang, L. R., Wong, L. and Goh, W. W. B. (2022b) ‘How doppelgänger effects in biomedical data confound machine learning’. *Drug Discov Today,* 27**,** 678-685.

Zhang, Q., Lu, J. and Jin, Y. (2021) ‘Artificial intelligence in recommender systems’. *Complex & Intelligent Systems,* 7**,** 439-457.