How to solve Doppelgängers effects

Introduction:

Doppelgänger effects refer to the phenomenon where two individuals who are not identical but appear to be similar, can cause confusion in various contexts. In machine learning, this effect can cause significant problems as it can confound the models that rely on distinguishing between these individuals.

For instance, imagine a facial recognition system that is trained to identify people from a dataset of images. If two people have very similar features, the system might not be able to distinguish between them accurately, leading to false identifications. This is where doppelgänger effects come into play.

In the case of facial recognition, doppelgänger effects can be caused by individuals who look similar, such as siblings, twins, or people from the same ethnic group. These individuals may have similar facial features, skin tone, and even hairstyles, which can make it challenging for the system to distinguish between them.

Another example of doppelgänger effects in machine learning can be seen in recommender systems. These systems are designed to recommend items to users based on their preferences and past behavior. However, if two users have very similar preferences and behaviors, the system may recommend the same items to both of them, leading to a lack of diversity in recommendations.

In both of these examples, doppelgänger effects can confound the models and lead to inaccurate results. To address this issue, researchers have proposed various methods such as feature engineering, data augmentation, and ensemble learning to improve the accuracy and robustness of machine learning models.

In conclusion, doppelgänger effects are a significant challenge in machine learning, as they can confound models and lead to inaccurate results. It is crucial to understand and address these effects to improve the accuracy and fairness of machine learning models in various contexts.

Doppelgängers effects in different types of datasets:

This effect does not only occur in medical datasets, it can also occur in different types of data.

Text data:

Doppelgänger effects can occur in text data when two documents or pieces of text have similar wording or language usage. This can lead to misclassification or confusion in natural language processing tasks such as sentiment analysis or topic modeling.

Environmental data:

Environmental data, such as air quality measurements or weather data, can be affected by Doppelgänger effects when two locations have similar environmental conditions. For instance, two cities may have similar weather patterns or air pollution levels due to their geographical proximity or other factors, leading to confusion in environmental monitoring and prediction.

Financial data:

Financial data, such as stock prices or transaction data, can be affected by Doppelgänger effects when two companies or investments have similar performance or behavior. This can lead to misclassification or confusion in financial modeling and decision-making.

Social network data:

Social network data, such as social media posts or friendship networks, can be affected by Doppelgänger effects when two individuals have similar social connections or post similar content. This can lead to misclassification or confusion in social network analysis and recommendation systems.

The occurrence of Doppelgängers effects in different types of medical data:

Genetic data:

Doppelgänger effects can occur in genetic data when two individuals have similar genetic profiles or family histories. This can lead to misclassification or confusion in genetic testing, diagnosis, and treatment.

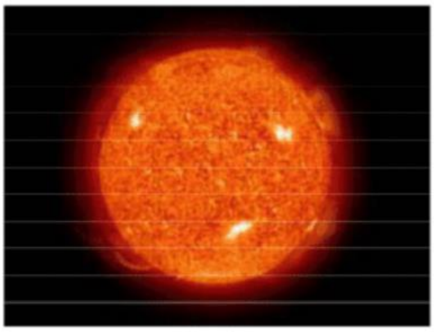
Electronic health records (EHR)：

EHR data can be affected by Doppelgänger effects when two patients have similar medical histories or symptoms. This can lead to misdiagnosis or incorrect treatment recommendations, especially for rare or complex diseases.

My current research project uses this type of data. In the process of processing the data, we use NLP to transform these Electronic health records (EHR) into numerical vectors. However, the lack of accuracy in NLP model may lead to some missing data and errors. Therefore, the final data is somewhat biased and may contain mislabelled data which would result in Doppelgänger effects.

Medical imaging data:

Medical imaging data, such as MRI or CT scans, can be affected by Doppelgänger effects when two patients have similar anatomical structures or imaging features. This can lead to misinterpretation or incorrect diagnosis in radiology and medical imaging.



The example given here shows the peach on the left and the sun on the right. However, the two images are very similar and can easily produce misleading results in the machine learning process

Omics data:

Omics data, such as proteomics or metabolomics data, can be affected by Doppelgänger effects when two samples have similar biomarker profiles or expression patterns. This can lead to misclassification or incorrect identification of disease subtypes or biomarkers.

How to identify of data Doppelgängers:

"Data doppelgängers" refer to instances in which two or more datasets exhibit a high degree of similarity or identity, potentially resulting from various sources, such as errors in data entry, inadvertent or intentional data duplication, or intentional data manipulation. Such occurrences can have significant implications for data quality, statistical analyses, and decision-making processes, necessitating rigorous data management and verification practices in research and data-driven applications.

First, we can use checksum algorithms to calculate the checksum for each data set. If the checksums are identical, it is likely that the data sets are identical.

For communication data, we can use the CRC (Cyclic Redundancy Check) algorithm, a widely used error detection technique for error control in data communication. It generates a checksum by dividing a block of data by a polynomial to detect if the data has been corrupted or tampered with.

For some medical data, MD5 (Message Digest Algorithm 5) can be used: MD5 is a widely used hash function that takes a message of arbitrary length as input and produces a 128-bit message digest as output. Due to its high degree of irreversibility and the fact that different inputs produce different outputs, MD5 is commonly used in applications such as digital signatures and message authentication.

The flow of MD-5 algorithm:

1. Initialize a message digest buffer to the initial value specified in the MD5 specification.
2. Pad the message so that its length is a multiple of 512 bits.
3. Divide the padded message into 512-bit blocks.
4. For each block, apply the MD5 compression function to the message digest buffer and the block.
5. After all blocks have been processed, the final message digest is the hash value.

Statistically, we can also use methods such as the pearson correlation coefficient and the spearman coefficient to calculate whether the data has Doppelgängers effects. In summary, data doppelgängers can be identified by comparing data sets, using checksum algorithms, using data deduplication tools, and using data visualization tools.

To Improve the doppelgängers effect in ML model:

When we find the doppelgängers effect, we can use the following methods to solve it：

Feature selection：

One of the main causes of the doppelgänger effect is the presence of redundant or irrelevant features in the dataset. Therefore, one effective strategy to improve the doppelgänger effect is to perform feature selection to identify and remove these features. This can help to improve the accuracy and interpretability of the ML model.

Model evaluation：

Model evaluation is a critical step in developing ML models that are robust to the doppelgänger effect. Cross-validation and bootstrap methods can be used to evaluate the performance of the model on different subsets of the data and estimate the generalization performance of the model. In addition, model interpretability techniques, such as feature importance analysis and decision tree visualization, can be used to understand how the model is making predictions and identify potential sources of the doppelgänger effect.

Ensemble methods：

Ensemble methods, such as bagging and boosting, can be used to improve the performance and robustness of ML models by combining multiple models that have been trained on different subsets of the data. This can help to reduce the impact of the doppelgänger effect by incorporating different sources of information and reducing the influence of individual samples that are similar to each other.

Conclusion:

The doppelgängers effect can not only make the model less credible, but also disturb our predictions. It can occur in any type of dataset. We need to use checksum method flexibly to identify it. Finally, we can use some methods to enhance our machine learning models to eliminate this effect.

Reference:

[1] Wang, L. R., Wong, L., & Goh, W. W. B. (2022). How doppelgä ngers effects in biomedical data confound machine learning. Drug discovery today, 27(3), 678–685. https://doi.org/10.1016/j.drudis.2021.10.017

[2] Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys (CSUR), 46(4), 44.

[3] Hassanien, A. E., & Zomaya, A. Y. (2018). Handling concept drift and class imbalance in data streams. In Big Data Analytics (pp. 69-92). Springer, Cham.

[4] Bifet, A., & Gama, J. (2010). Adaptive learning from evolving data streams. In Machine Learning and Knowledge Discovery in Databases (pp. 19-46). Springer, Berlin, Heidelberg.