1:

Here is the name of the paper I want to share.

2:

3:

Although machine learning is deeply applied in materials science, the existing datasets have notable imbalances. And most of the works focus on feature extraction and ML algorithms, and only a few people consider the effect of the material dataset.

In this article, the author focuses on the glass-forming ability of ternary alloy. They found that when using the original dataset, there were some overfitting issues. They try to use the balanced dataset to mitigate it.

4:

For the original dataset, all data are from two kinds of experimental methods: high-throughput and traditional experiments. The high-throughput method would have abundant, low-fidelity high-throughput data. And traditional would produce less data with high fidelity. There are 5 alloy data from high-throughput and 12 from traditional experiments.

For the balanced dataset, the author reduced the data from high-throughput experiments and increased the data from traditional experiments. They reduced the number of data from each high-throughput experiment to 200. And augment the traditional experiment data by increasing data from the existing compositions with a step of 0.1 percent.

The balanced data is smoother.

5:

The article uses an ANN model. It has an input layer with 131 features. And two hidden layers. And one output layer.

They trained the model with original and balanced data.

6:

And from the table, balanced data is more accurate.

We can find the root mean square error of balanced data is lower, especially for the traditional data part. It is overfitting in the ANN-original model.

7:

The author also uses leave-one-alloy-system-out validation to test the ability to predict a new alloy. They leave the data in one alloy system completely out of the training dataset, and such an alloy system only serves as the validation dataset. For example, aluminum-magnesium alloy. Related data is from traditional experiments. Ann-original data is not very accurate and it keeps low glass-forming ability for the whole region. But ANN-balanced mode; can learn more information from the train data and get a better prediction.

8:

This method of balancing the dataset can improve the accuracy of the predictions of GFA and it is very useful for people to improve the quality of the dataset in the future.