Machine Learning Final Project Report

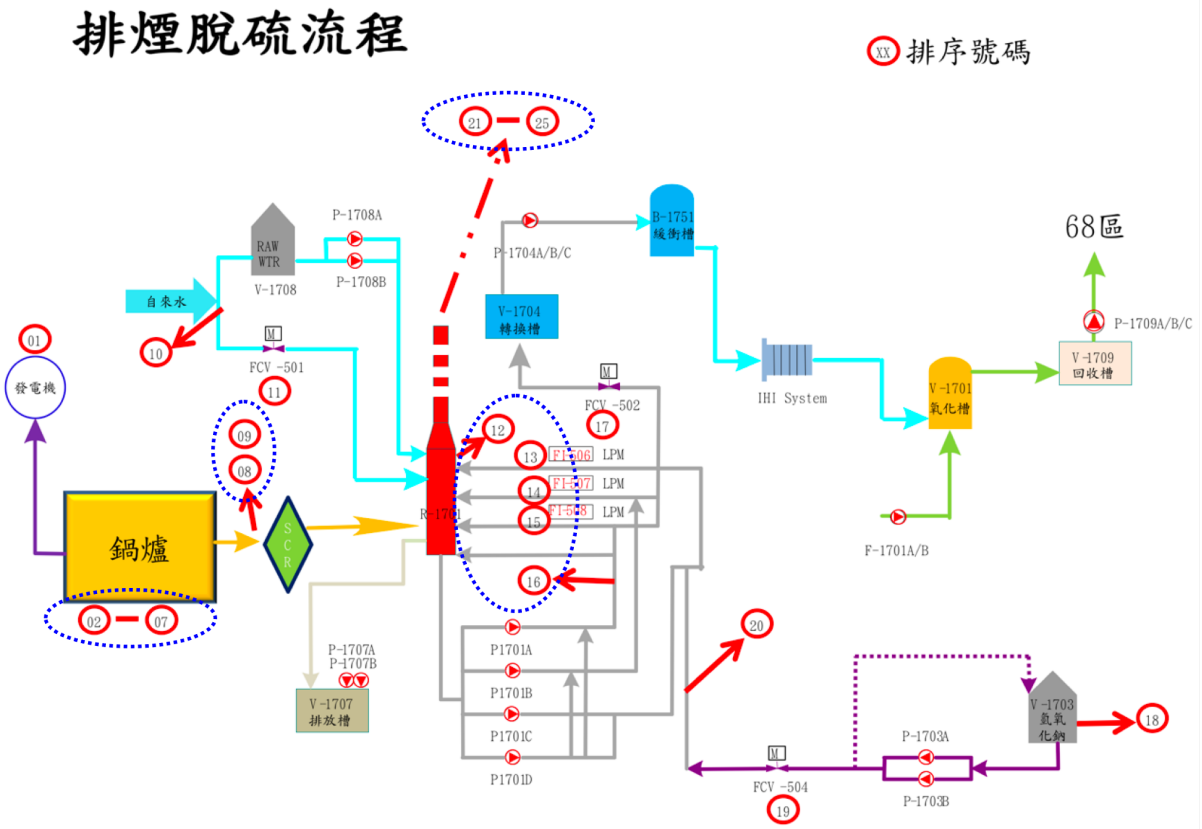
Team 32 盧冠維 109061621 劉軒宏 101061620

Introduction

In recent years, the industry has developed rapidly, but a large amount of waste gas has been emitted while developing the industry. Among these exhaust gases, the proportion of sulfur dioxide is quite high. Therefore, we need a stable and effective desulfurization system for exhaust gas desulfurization to reduce environmental pollution. The training data and test data we use come from different parts of the desulfurization system equipment. A total of 25 equipment detection variable records are provided. We use these 25 test variables to assess the pollution level, and the goal is to classify 5 different levels of pollution standards, to find out the correlation between different parts of the desulfurization system and the pollution level.

Approach

In this experiment, we apply different models such as linear and non-linear models to compare the accuracy of pollution level prediction. Trying to analyze the characteristics of each model, and add the concept of data grouping, e.g. PCA, as shown in the blue dashed box in the figure below, trains the data in the same area together, see whether it can improve the accuracy of prediction.



Model

The linear models we applied in this experiment are Linear Regression, K Nearest Neighborhood, Decision Tree, and Random Forest, for the non-linear model, we applied a 1-dimensional Convolution Neural Network. Further, after the demo, we also try to apply Linear Discriminant Analysis as one of our evaluations.

Experiment

We use the models listed above to fit the data and try to predict the level of pollution. The important part is that we encode the date and time as one of the features because we found that the pollution level is related to the time before and after, which means the date and time have to be taken into consideration. We also use PCA to preprocess some data that come from similar parts.

Result

Besides the 24 features that are given in the data, we encode the date and time as an additional feature, making our input features with 25 features. As shown in figure 1 and 2, you can tell that while adding the date encoding, the performance has been improved no matter which model it is. Though the improvement is slight, it confirms that the pollution level is related to the time before and after, and it's important to take date and time into consideration.

Fig. 1, Loss of all models

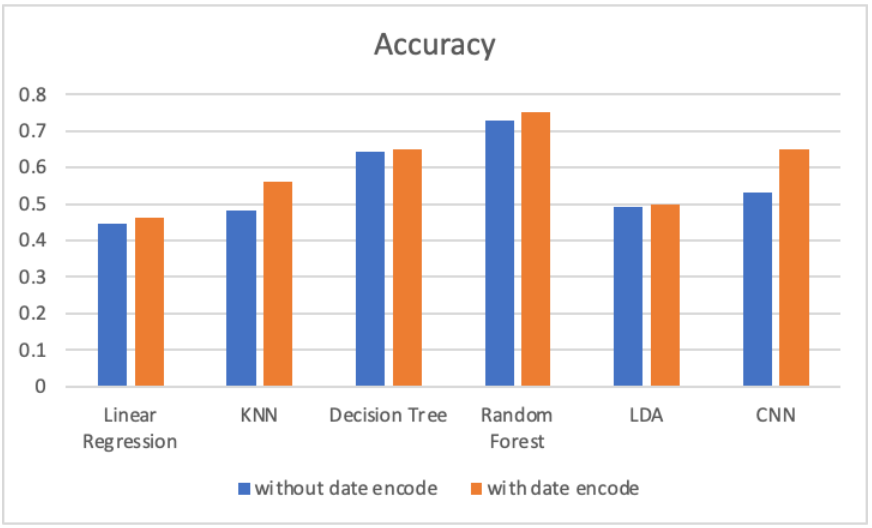


Fig. 2, Accuracy of all models

Furthermore, we found that the class distribution of the training data is unbalanced, so we applied weighted softmax to our Convolution Neural Network model, to prevent it from overfitting specific classes and underfit to others at the same time, resulting in a performance drop. We also provide the result of with and without weighted softmax in figure 4 and 5. As you can see, the results of without weighted softmax are better than adding weighted softmax, that is because the model is overfitted to class 1 and 2 but still underfit to class 4 and 5, so when it comes to testing set, the result of adding weighted softmax is better.

We also report the result of using PCA to the features that come from a similar part of the desulfurization system in figure 4 and 5. It turns out that using PCA to similar features is useless, the reason is that the features after PCA are concentrated, PCA didn't help very much.

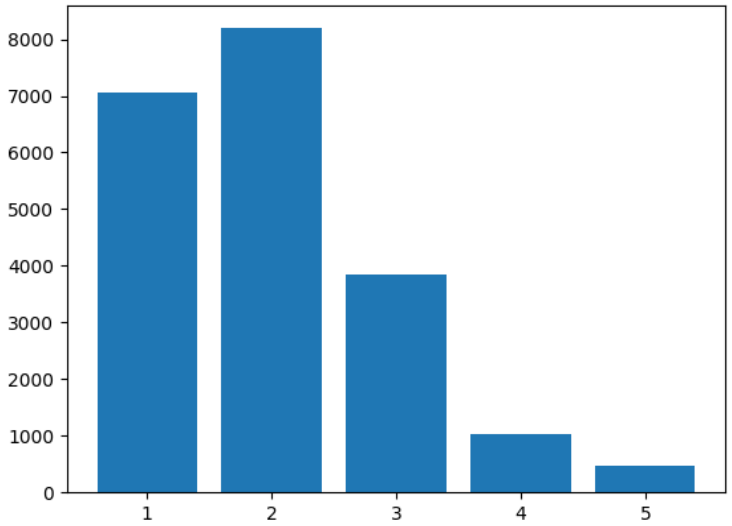


Fig. 3, Distribution of the data classes

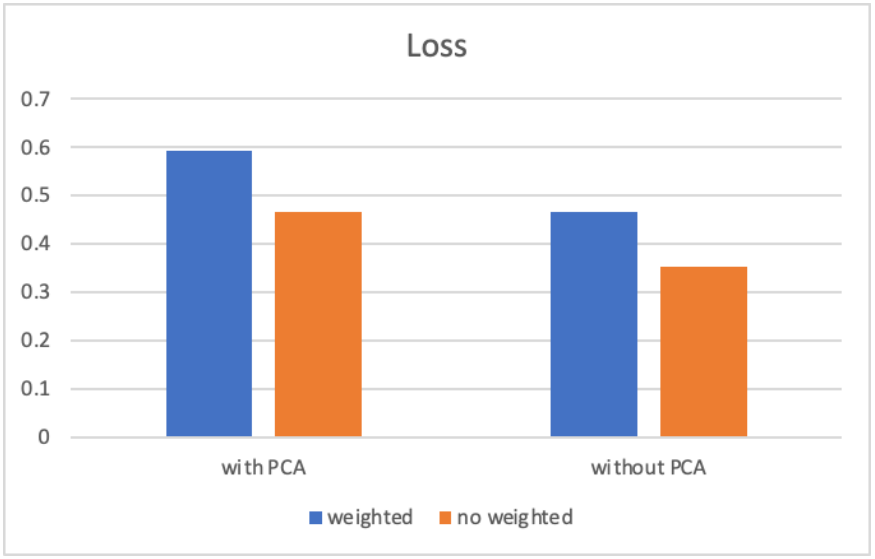


Fig. 4, Loss of with and without PCA and weighted softmax

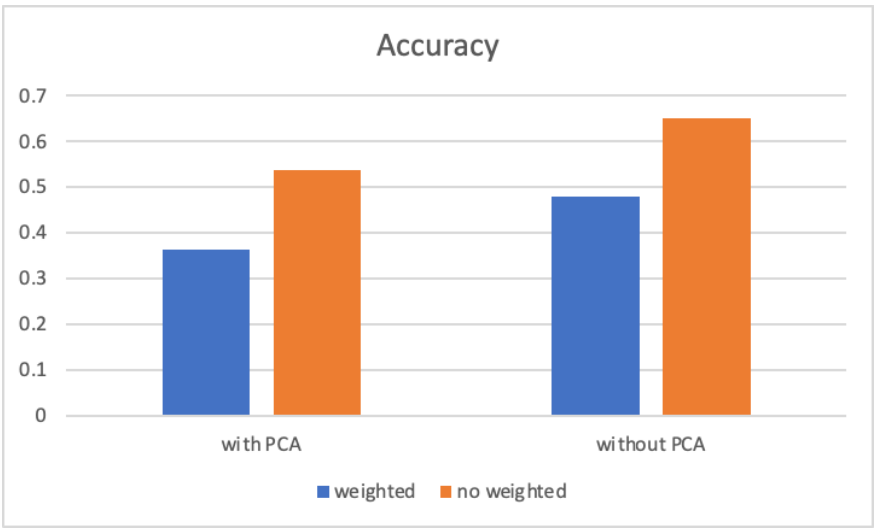


Fig. 5, Accuracy of with and without PCA and weighted softmax

Conclusion

After the experiment, we find that time and date are important features in this task because the pollution level is related to the time before and after. We assume that apply RNN to this task might achieve better performance because RNN can utilize the information of time much better than CNN. PCA and LDA didn't help to improve the performance in this task because the features are very concentrated after dimension reduction. And, adding weighted softmax to address the unbalance in the training data is crucial, because it affects the performance a lot.

Reference

1. 課堂講義 Chapter 2 : Linear Regression, K-Nearest Neighbor

2. 課堂講義 Chapter 5 : Convolutional Neural Network
3. Wikipedia : Decision Tree,Random Forest