MLCLASS_FDU Report for CSCI567 Project

Junming Chen

Zekun Li
Affiliation
Address
email

Junfeng Wu
Affiliation
Address
email

Abstract

This is a project using Random Forest Algorithm to predict hourly rainfall. After data preprocessing, the optimization is mostly done by feature transformation and selection. The final result reached 23.75398 variance and rank 51 of all teams.

1 Data Processing

1.1 Overview

First of all, there are about 38% totally missing data in the data set and many data entries with incomplete data. As the MAE changing of this competition, we only predict one target value for each ID's data and process those data with at least one of the Ref values is non-null.

1.2 Preprocessing

For the data preprocessing, the first rule is to extract only valid data entries, which has at least one of the Ref values is non-null. Then according to common sense, we eliminate the reflectivity values, composite values and reference values which are below zero.

It is also essential to eliminate abnormal data and some outliers. We firstly eliminate the negative values in features Ref, RefComposite because they should be positive. From the expected rain frequency histogram, we found out that, most values are lower than 100. However, there are some extremely large values more than 1,000 and should be eliminate. We choose the threshold of 70 at last because the bucket element droped to only 300 from bucket (68,70) while all the bucket in range (0,68) with interval of 2 have more than 2,000 data entries fall into them.

2 Feature Selection

2.1 Basic matrice

From the official data description, we have to generate only one line of feature value for each given data group by ID. Thus, we firstly transfer the given data using mean, variance or summation.

We first plot some density graph and saw that the mean value of Ref seems correlated with expected value. The variance of Ref is different from mean but we can see from the plot that most

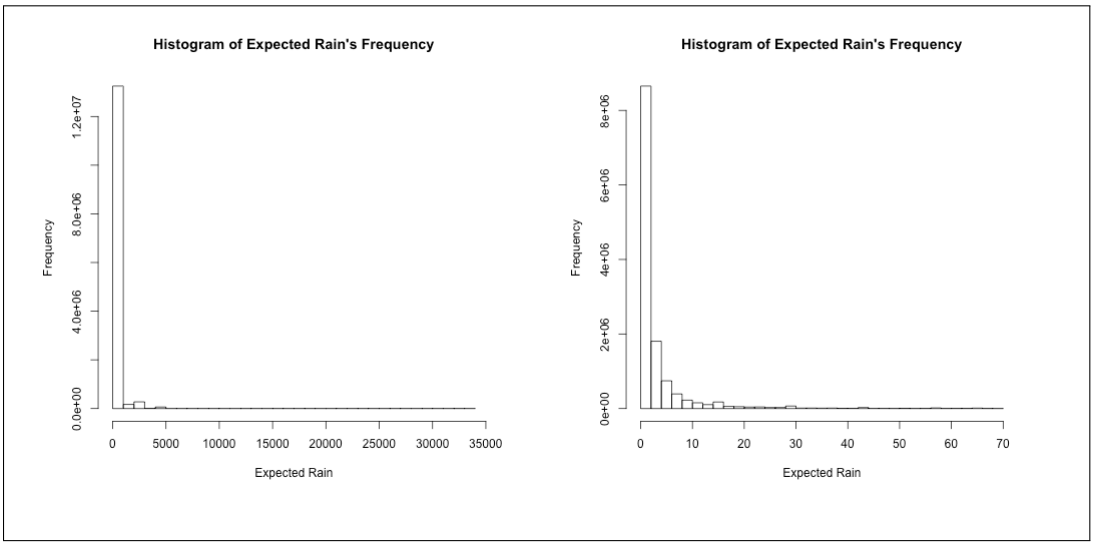

Figure 1: Density scatter of relation between Ref, RefComp50, Ref50 and Expect Rain

variance is relatively low with expected value lower than 2. Thus, using these features could be helpful for the prediction.

Some basic features we selected:

- **Ref**: Using mean, variance and summation of Ref group by ID.
- **Ref_5x5, RefComposite and RefComposite_5x5**: Using both mean and variance of each feature group by ID.
- **rhoHV, RhoHV_5x5, Zdr and Zdr_5x5**: Using only mean of each feature group by ID.
- **rhoHV, RhoHV_5x5, Zdr and Zdr_5x5**: Using only mean of each feature group by ID.
- **Kdp**: Using only mean of each feature group by ID. This didn't work well so not used in final features.
- **Number of records**: Using the record number in an ID group as a feature.
- **Number of Ref's missing values**: Using the number of Ref's missing value in an ID group as a feature.

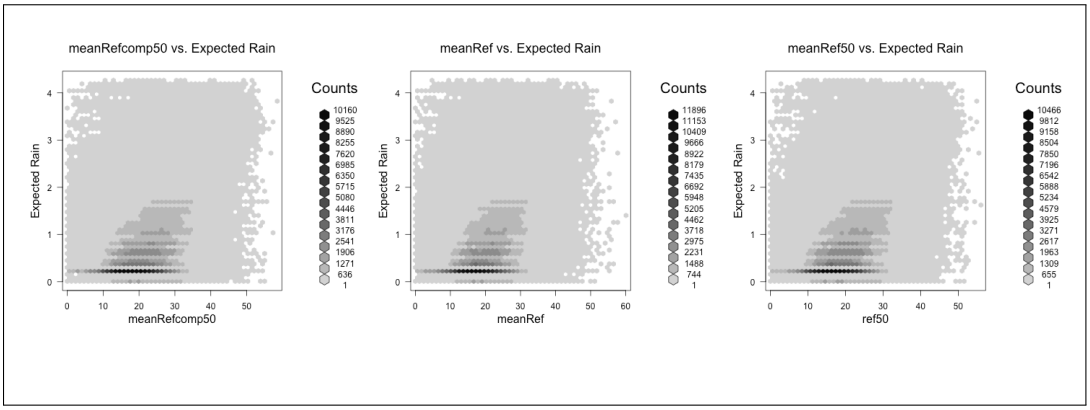

Figure 2: Density scatter of relation between mean Ref, RefComp50, Ref50 and Expect Rain

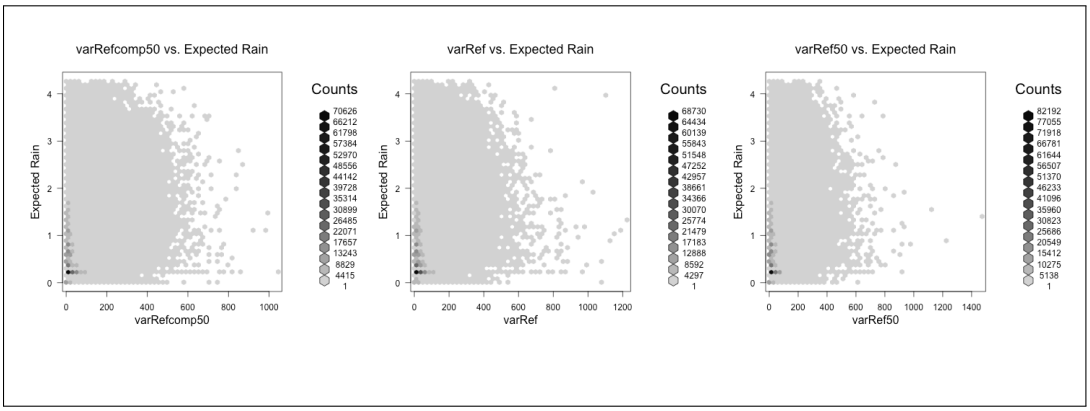

Figure 3: Density scatter of relation between variance Ref, RefComp50, Ref50 and Expect Rain

2.2 New Features

Combining Marshall-Palmer Observation: From the official description, Marshall-Palmer relationship is a method to predict the gauge observation using radar data. It used the Ref data combined with minutes_past to calculate the prediction and it is a reasonable way to take minutes_past into account.

We simply add the Marshall-Palmer observation into the training feature, but it will lower the prediction accuracy so not used in final features.

Combining transferred KDP: Because the former combination failed, we sought another way to do numeric calculation of predicted raining rate. We found a new formula which can also do this.

$$rate = \sum((sign(Kdp) * (kdpzdr\_aa) * (|Kdp|^{kdpzdr\_bb})(ZDR^{kdpzdr\_cc})) \quad (1)$$

where $kdpzdr\_aa = 136, kdpzdr\_bb = 0.968, kdpzdr\_cc = -2.86$.

We added the raining rate predicted by KDP into the training feature, but it will lower the prediction accuracy so not used in final features.

Combining radardist_km: Because radar usually cover a certain area which is related to the radardist_km. Then we examined new features of original Ref and RefComposite divided by radardist_km, and original Ref and RefComposite divided by radardist_km^2.

The second new feature using radardist_km^2 seems more reasonable because if the cover area of the radar is a circle, the area is related to radardist_km^2. However, the cross validation result showed that the first new feature works better, so we only added original Ref and RefComposite divided by radardist_km as for final features.

Combining difference on Ref between two consecutive time: Because we haven't got used of time, we tried to add features of averaged Ref and RefComposite difference between two consecutive record in same ID. The seemingly more reasonable feature is to add a feature which is the consecutive Ref and RefComposite difference divided by time difference. However, It worked worse than the former one. So only averaged Ref and RefComposite difference was added into the final features.

2.3 Final Features

We do the feature selection process and only add a feature when adding it can improve the cross validation's mean squared error. Finally 40 features are selected and can be categorized as:

- **Ref(including all percentile data)**: Using mean, variance and summation.

- **RefComposite(including all percentile data)**: Using mean, variance.
- **rhoHV and zdr(including all percentile data)**: Using mean.
- **mean Ref and RefComposite divide radardist_km**: Using value of this equation.
- **difference on Ref and RefComposite between two consecutive time**: Using value of this equation.

3 Algorithm

Our first attempt is using Random Forest Algorithm. The reason is:

- It is good classifier for most data and scenes.
- It can handle large amount of variables.
- It gives features different importance.
- It can work even with high percentage missing data.
- The speed is relatively fast.

4 Evaluation

4.1 Evaluation on Feature Selections

We implement the Random Forest regression based on the starter script in Kaggle Forum. In the starter script, the selected features are:

- mean and sum of Ref
- mean of Refcomp
- Distance
- Number of records and missing values

Table 1 shows the improvement after adding the features we selected. And Table 2 shows the importance among all the selected features

Table 1: Feature Selection Evaluation

	MAE
Random Froest in starter script	23.89669
Random Froest after adding features	23.75398

4.2 Evaluation on Different Regression Method

References

References follow the acknowledgments. Use unnumbered third level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to 'small' (9-point) when listing the references. **Remember that this year you can use a ninth page as long as it contains *only* cited references.**

[1] https://www.eol.ucar.edu/projects/dynamo/spol/parameters/rain_rate/rain_rates.html

[2] http://glossary.ametsoc.org/wiki/Marshall-palmer_relation

Table 2: Variable Importances

Order	Variable	Scaled Importance	Percentage
1	mean of Ref_5x5_90th	1.000000	0.148810
2	mean of Ref_5x5_50th	0.631968	0.094043
3	mean of RefComposite_5x5_90th	0.501650	0.074651
4	mean of Ref	0.421761	0.062762
5	Radardist_km	0.281814	0.041937
...			
41	max of Kdp	0.063436	0.009042
42	mean of RhoHV_5x5_90th	0.062761	0.008945
43	min of Kdp	0.062702	0.008937
44	mean of RhoHV_5x5_10th	0.059808	0.008524
45	mean of RhoHV_5x5_50th	0.059564	0.008490

Table 3: Feature Selection Evaluation

Model	MAE
xxx	23.79669
Random Froest	23.75398