Final Report

rlz818 yrk296 yxe836

Introduction:

Our project is a radar-based Quantitative Precipitation Estimates (QPE) problem, estimating the hourly rainfall by using data collected from the radar. In our daily life, it's very important to know how much it rained on a particular field, because rainfall will affect our society and lives largely, especially for agriculture. However, rainfall is variable in space and time, and it happens suddenly, so it's hard and expensive to use rain gauges everywhere to estimate the rainfall. Therefore our project can help to estimate the rainfall via radar observations, which is cheap and feasible.

Data:

Our data is collected from the radar, and it contains 24 attributes about:

Ref(Radar reflectivity in km)

RefComposite(Maximum reflectivity in the vertical column above gauge)

RhoHV(Correlation coefficient)

Zdr(Differential reflectivity)

Kdp(Specific differential phase)

The most important attribute is Ref, which is highly related to the rainfall. Our raw data set contains more than 1000000 ids, and each id represents a radar's observation in 6 time slots. Also, in each observation, we have an expected rainfall value, which is collected by the rain gauges and it represents the real rainfall in those area in that time slot. Because our data is too big and it contains many blank or outliers, so we have to do preprocessing before training to discard those abnormal data. Finally we will use ten-fold cross validation to output the MAE(Mean Absolute Difference) as our results.

Method:

In our project, we use two methods to estimate the rainfall: Random Forest, XGboost and IBk(k-nearest neighbors)

1. In Random Forest, we wrote a program using R library to work out the results.

We firstly choose the most important attributes Ref(including Ref_5*5_10th, Ref_5*5_50th, Ref_5*5_90th, the number 10, 50 and 90 means 10th, 50th, and 90th percentile of reflectivity values in 5*5 neighborhood around the gauge) for training to come up with a result(our results have ten values because we use 10-fold cross validation):

	1	2	3	4	5	6	7	8	9	10
MAE	2.7060	2.8892	2.5275	2.7342	2.6804	2.5695	2.6098	2.8625	2.5728	2.5899

Ref

Moreover, we try to add more secondary attributes on it for training, like Zdr(including Zdr_5*5_10th, Zdr_5*5_50th, Zdr_5*5_90th) and Kdp(including Kdp_5*5_10th, Kdp_5*5_50th, Kdp_5*5_90th), and yield another result for comparison:

	1	2	3	4	5	6	7	8	9	10
MAE	2.6878	2.8776	2.5178	2.7277	2.6789	2.5626	2.5829	2.8928	2.5958	2.5669

Ref + Zdr + Kdp

We also use a dBZ meteorology(stands for decibel relative to Z, which is a logarithmic dimensionless technical unit used in radar, mostly in weather radar, to compare the equivalent reflectivity(Z) of a radar signal to the return of a droplet of rain with a diameter of 1mm to estimate the rain intensity) to calculate the rainfall rates and combine it with our training to generate a more accurate result.

In order to have a lower MAE, we tried many proportion of combination. For example, we use 0.5 dBZ + 0.5 RZK(including Ref, Zdr, Kdp) for training and obtain a result:

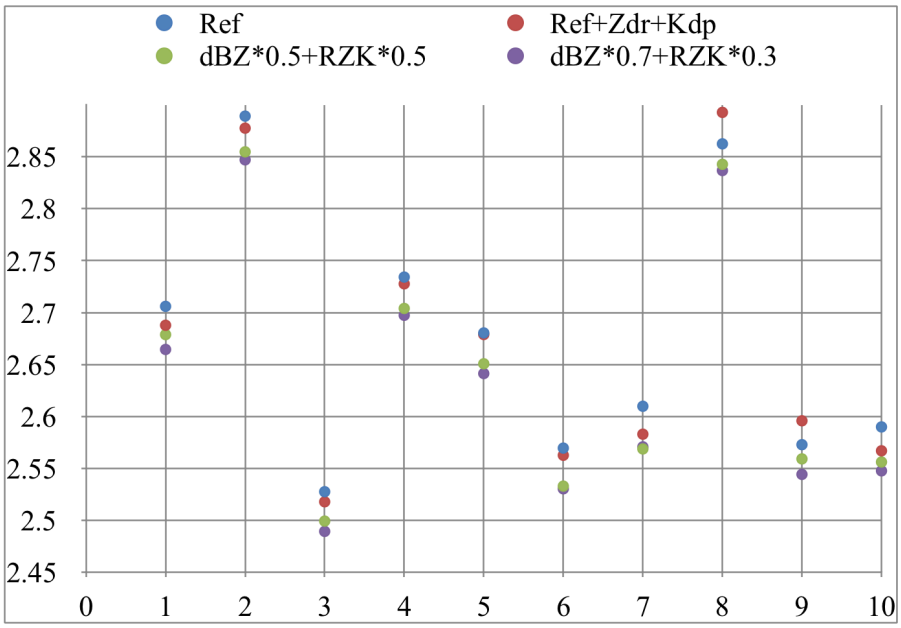
	1	2	3	4	5	6	7	8	9	10
MAE	2.6788	2.8548	2.4992	2.704	2.6507	2.5329	2.5687	2.8428	2.5591	2.5561

0.5 dBZ + 0.5 RZK

In the combination of 0.7 dBZ + 0.3 RZK, we gain a lowest MAE:

	1	2	3	4	5	6	7	8	9	10
MAE	2.6645	2.8471	2.4893	2.6974	2.6413	2.5303	2.5707	2.8368	2.5442	2.5476

0.7 dBZ + 0.3 RZK



comparison of different implementations of Random Forest

2. XGBoost is used for supervised learning problems, where we can use the training data x to predict a target variable y. In our XGBoost, we wrote a program based on R library, using attributes Ref(including Ref_5*5_10th, Ref_5*5_50th, Ref_5*5_90th) and RefComposite as training data x, because they are the most important attributes, to yield the result of MAE:

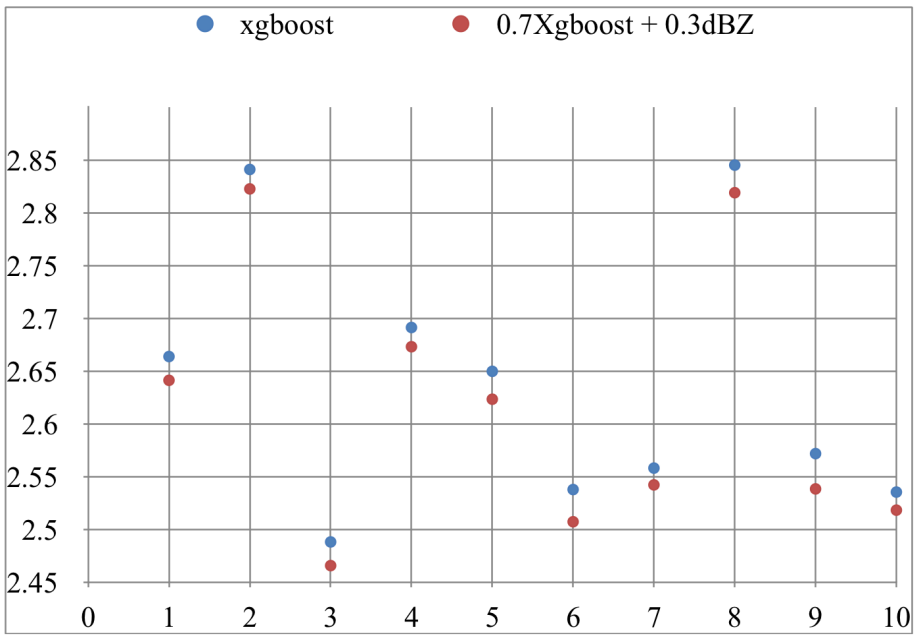
	1	2	3	4	5	6	7	8	9	10
MAE	2.6639	2.8410	2.4884	2.6914	2.6499	2.5380	2.5583	2.8451	2.5720	2.5356

XGBoost

We also tried to add the dBZ meteorology to XGBoost to obtain a lower MAE:

	1	2	3	4	5	6	7	8	9	10
MAE	2.6414	2.8226	2.4660	2.6732	2.6235	2.5075	2.5424	2.8190	2.5386	2.5185

0.7 XGBoost + 0.3 dBZ



compariaon of different implementations of XGBoost

3. In IBk(k-nearest neighbors), we put our data into Weka and use it to work out another result:

	1	2	3	4	5	6	7	8	9	10
MAE	2.0605	2.1721	2.0265	2.0624	2.1378	1.9871	1.9088	2.0753	2.1201	1.9650

IBk

In conclusion, the MAE in IBk < XGBoost < Random Forest, that means IBk is best. But XGBoost may have the largest potential to come up with a best result.

Future Work:

Because XGBoost needs much time for calculation, we don't have enough time to implement it with all the attributes. Also, in IBk, we use one tenth of the data(about 100000) to work out the final result due the limit of time. Therefore, in the future, we can try to combine the dBZ meteorology with XGBoost or IBk and train all our data with all the attributes to come up with a better result.

Work Division:

Rui Liao works on the data processing and the final report; Yiyi Ren works on

training data and MAE calculation; Ye Xue works on building the webpage.

Thanks:

In this quarter, we have spent a great time studying with Professor Downey and learned a lot of machine learning, which is powerful and glamorous. The class is fantastic and useful, and we hope we can use what we learned from it to cope with other more challenging problems in the future.

Thank you very much and wish you a nice summer!