A Method for Short-Term Quantitative Precipitation Forecasting

Zhi Zhang
Xi'an Jiaotong University
Xi'an 710049
China
zhangzhi@stu.xjtu.edu.cn

Shenghua Wei
Xi'an Jiaotong University
Xi'an 710049
China
weishenghua@stu.xjtu.edu.cn

ABSTRACT

Short-Term Precipitation Forecasting is a task to predict a short-term rainfall amount based on current observations, which is a very important problem in the field of meteorological service. The radar echo extrapolation data is always used to predict the rainfall amount. In our method, the radar echo extrapolation data is regarded as a 2-D map. We propose a method based on the histogram of intensity (HOI) of the radar maps at different time point and different height. The residual of HOI of consecutive radar maps are computed to represent the trend of the radar map in time domain. The Principle Component Analysis (PCA) is used to extract main influencing factors from the original HOIs and residuals. Our method achieves the RMSE of 12.33323 on the dataset provided by CIKM 2017, which consists of radar map information and the rainfall amounts of different sites collected by meteorological observatory centers.

KEYWORDS

Histogram of Intensity (HOI), Residual, Principle Component Analysis (PCA), Radar Map

1 INTRODUCTION

The weather prediction and forecast is a critical meteorological service, which supports casual usages such as outdoor activity and even provide early warnings of floods or traffic accidents.

Short-term precipitation forecasting such as rainfall prediction, one type of the weather predictions, is a task to predict a short-term rainfall amount based on current observations. The radar data, rain gauge data and numerical weather are always used to predict the short-time rainfall amount. In the task of the CIKM2017, the radar data is used to predict the rainfall amount in the future.

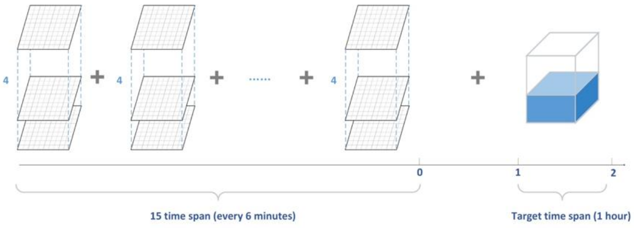

Figure 1: The illustration of the dataset and the mission. 15 time spans of radar echo extrapolation data at 4 different heights are provided to predict the rainfall amount of the target time span (future 1-hour to future 2-hour).

Radar is an object-detection system that uses radio waves to determine the range, angle, or velocity of objects. The radar echo extrapolation data is the radar reflectivity of the cloud of different time at different height. The radar data is represented by a 2-D array, and also referred as a 2-D map. As shown in Fig. 1, there are 15 time spans of radar maps before the current time. At each time span, there are 4 radar maps at each time points at 4 different height. The target is to use the radar maps to predict the rainfall amount in the future 1 hour to 2 hour.

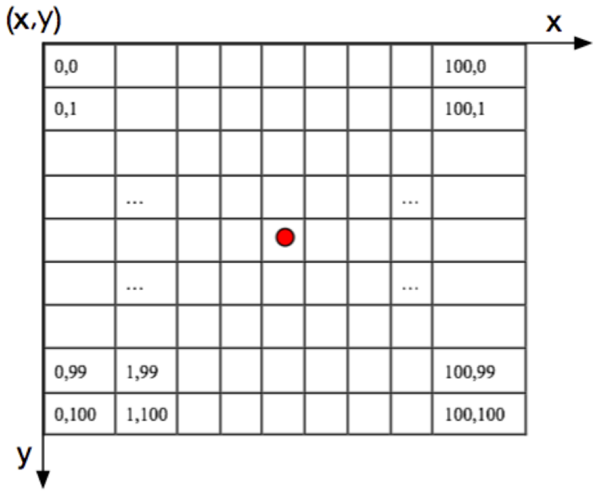

Figure 2: The illustration of the radar map.

As shown in Fig.2, each radar map covers a target site and its surrounding areas. It is marked as a m*m grids where each grid point has a radar reflectivity value z. The value of Z is measured by dBZ and linear transformed for anonymization purpose. The target site which rainfall amount should be predicted is the center of the radar map. Every grid of the radar map represents a 1 km* 1km site. Each radar map covers an area of 101*101km^2.

The shape of the radar maps and the variation and trend of consecutive radar maps should be used to represent the attribute of the weather of a specific site.

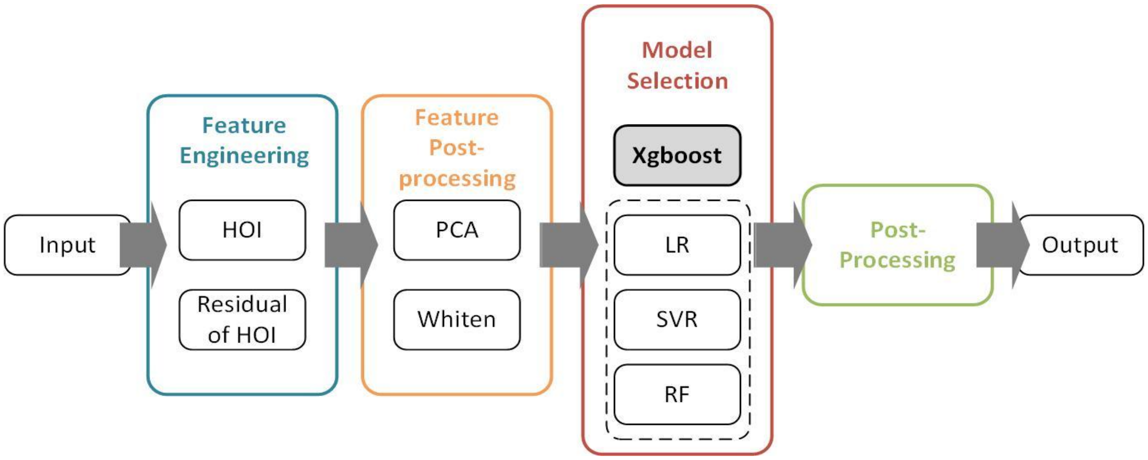

Figure 3: The work flow of our method.

We propose a method based on histogram of intensity (HOI), principle component analysis and Xgboost, as shown in Fig. 3. Firstly, we compute the HOI of each radar map. Then the residual of consecutive HOI is computed to represent the trend of the time series. The HOI and its residual form the original feature vector with high dimensionality, which makes the regressor hard to train. To deal with this problem, we use the principle component analysis with whiten operation to extract the main influencing factors of the original feature vector. Finally, we do post-processing by eliminating the negative numbers in the predicted rainfall amount. Our method achieves the RMSE of 12.33323 on the CIKM2017 dataset.

2 Related Work

Tracking Radar Echo by Correlations (TREC) technique [1] is always used to estimate the motion field of the radar map. For two consecutive radar maps of the same site at time t and t+1, they are firstly separated as small cells. For every cell in the radar map at t, the correlation coefficent between itself and the potential matching cell in the radar map at t+1. The cell in radar map t+1 with the largest correlation coefficent is regarded as the matched cell. The motion vector to move from the cell at time t to the matching cell at time t+1 is the motion of radar map at the center of that cell. The motion vectors of the whole radar map represents the trend of the radar maps. Besides, there are more similar methods like TREC, such as COTREC and DITREC. However, the computation of correlation coefficents of cells consume a large time. Sometimes the motion estimation is not accurate, which influences the prediction of the rainfall amount.

The Histogram of Gradients (HoG) [2] is always used to describe the appearance and shape of images. But the HoG feature does not fit the radar data. The local texture of small cell varies from each other and does not directly influence the rainfall amount. The intensity distribution of cell present a coarse-grained attribute of an area and is robust to noise, so we use the histogram the intensity of the radar map to represent the weather attribute.

Principal component analysis (PCA)[3] is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation). PCA is mostly used as a tool in exploratory data analysis and for making predictive models. PCA is used in our method to reduce the feature dimensionality.

XGBoost[4] is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. The xgboost is always used both for classification and regression. We use the xgboost to predict the rainfall amount, together with the feature extracted and transformed by PCA.

3 Method

3.1 Feature Extraction

The shape and the texture of the radar maps describe the shape and attribute of cloud in the sky. The baseline method provided by the CIKM2017 use all the intensity value of a radar map to represent the cloud condition. Using raw radar map value to form the feature vector has two disadvantages: first, high dimension of the data makes training and testing fairly slow; besides, the raw data is not discriminative to represent cloud conditions. Therefore, we use Histogram of intensity to represent each radar map, which can denotes the intensity distribution of the local regions on the radar maps.

For each 101x101 radar map, we divide it to 9 cells by a 3x3 grid. In each cell, the normalized histogram of intensity is computed. All the HOIs have the intensity range from -1 to 215, which come from the statistics of all the pixels in the cells. There are 16 bins for each histogram. The HOIs are transformed as the density form because some radar maps have missing values. The total dimension of hoi for each sample is 9(cells per radar map) x 16(bins per cell) x 4(heights per time span) x 15(time spans) = 8640.

Table 1: Results on the CIKM 2017 dataset with HOI and different settings.

Settings	RMSE(CV)	RMSE (test1)	RMSE(test2)
HOI+xgboost	14.60	13.56	N/A
HOI+PCA+xgboost(PCA on train set and test set 1)	14.00	13.32	13.84
HOI+Residual+PCA+xgboost(PCA on train set and test set 1)	13.63	N/A	13.23
HOI+Residual+PCA+whiten+xgboost(PCA on train set and test set 1)	13.31	N/A	13.03
HOI+Residual+PCA+whiten+xgboost(PCA on train set, test set 1 and test set 2)	13.23	N/A	**12.33**

Table 2: Results on the CIKM 2017 dataset with different features.

Settings	RMSE(test1)
LR+ALL map center	14.79
Xgboost+Motion hist	14.68
Xgboost+HOG	14.44
Xgboost+HOI	13.56

The HOIs can describe the intensity distribution of single radar map at current time. In order to describe the changing trend of the radar maps in the time domain, we also compute the residual of consecutive radar maps. The HOI feature of time t can denotes as H_t . For time t and time t+1, the residual is computed by subtracting the HOIs of time t+1 from the HOIs of time t, which can be computed as :

$$R_t = H_{t+1} - H_t \text{ , } t \in \{1,2,\dots,14\}$$

The dimensionality of the HOIs of time t is 9(cells per radar map) x 16(bins per cell) x 4(heights per time span) = 576. There are 14 residuals for 15 time spans and the total residual feature is a 8064-d (576x14) vector.

We combine the HOIs and its residuals to represent the raw radar map, which can not only describe the intensity distribution of radar map at current time but also the trend of intensity change. The original feature vector of a sample consists of the HOIs and its residuals. The original feature vector has 16704 dimensions (8640+8064).

3.2 Principle component analysis with whiten

After the feature is extracted, we use the principle component analysis to reduce the feature dimension. It has two reasons: on the one hand, the raw feature have a great many dimensions which make regression model difficult to predict rainfall amount; on the other hand, some dimensions in the raw feature have high correlation.

There are three parts of the CIKM2017 dataset: the train set, the test set 1 and the test set 2. We concatenated the feature vectors of the samples in all the three datasets. We get the first 3000 components with higher eigen values. Then we project the three datasets to the new space composed by the 3000 components.

Before PCA, we whitened the feature vectors to improve the predictive accuracy of the downstream PCA estimators by making their data respect some hard-wired assumptions. After the PCA with whiten, a feature vector is reduced to 3000d.

3.3 Model Selection and Training

After feature extraction and PCA dimension reduction, we need to choose suitable regressor with good hyper-parameters. We split the train set into 10 folds for cross validation. We tried the Xgboost, Logistic Regression (LR) [5], Support vector regression (SVR) [6] and the Random Forest (RF) [7] with different hyper-parameters. According to the cross validation results on the training dataset, Xgboost regressor with 300 estimators and max depth of 3 performs best. We use the Xgboost as the final regressor.

3.4 Post processing

Although in the training set, there is no negative values in the ground truth rainfalls, some negative values will be predicted in the test stage. We post processed the predicted rain amount by replacing all the negative values by zeros.

4 Experimental Results

There are 3 parts of the CIKM2017 dataset, the train set, the test set 1(stage 1) and the test set 2(stage 2). We do the model selection on the train set by using a 10-fold cv.

As shown in Table 2, in the first stage of the competition, we tested different regressors, including logistic regression (LR) and Xgboot, etc. We also extracted different features, such as the center value of each radar map, the motion vector and the motion histogram of radar maps estimated by TREC. We also tried the HOG feature of the radar maps. It turns out that the HOI is the best feature with the lowest RMSE on the test set 1.

As shown in Table 2, we tried different settings on the HOI feature to enhance the performance. We firstly test HOI with xgboost. After adding PCA on the train set and test set1, the RMSE reduced in the test set1. After adding the HOI residual into the feature, the RMSE reduced on the test set2. After adding the whiten operation, the RMSE reduced more on the test set2. When we do the PCA on all three parts of the dataset, the RMSE on test set2 achieves 12.33, which is the second lowest RMSE in the competition.

5 Conclusion

We propose a method for short-Term Quantitative Precipitation Forecasting. Firstly we calculate the HOIs of radar maps and their residuals. Then a PCA with whiten is operated on the original feature vectors. Finally, the xgboost is trained. The experimental results on the CIKM2017 datasets shows the superiority of our method.

ACKNOWLEDGMENTS

We would like to express our gratitude to all those who helped us during this competition, especially Mrs. Yonghong Song and Mr. Yuanlin Zhang.

REFERENCES

[1] Smythe G R, Zrnic D S. Correlation Analysis of Doppler Radar Data and Retrieval of the Horizontal Wind.[J]. Journal of Applied Meteorology, 2008, 22(2):297-311.

[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005:886-893.

[3] Shlens J. A Tutorial on Principal Component Analysis[J]. 2014, 51(3):219-226.

[4] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016:785-794.

[5] Jr D H W, Lemeshow S. Applied Logistic Regression[J]. Journal of the American Statistical Association, 1989, 34(1):358-359.

[6] Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2002, 2(1):45-66.

[7] Breiman L. Random Forest[J]. Machine Learning, 2001, 45:5-32.