Combined XGBoost Model for Short-term Percipitation Nowcasting

Report of Team 8 for CIKM AnalytiCup 2017

Anqi Liu
Telecom Paris-tech
anqi.lsun@gmail.com

ABSTRACT

In this paper, we present a XGBoost regression model for the weather prediction in CIKM AnalytiCup 2017. The main target of the competition is to use radar echo extrapolation data to do an accurate short-term rainfull prediction.

Considering the format and scale of given data (10000 samples in $15 \times 4 \times 101 \times 101$ format, 15 timesteps, 4 heights, $101 \times 101 km^2$ area), we adopt the XGBoost model to handle this problem. We exame various strategies of feature engineering and model tuning. Experiment results demonstrate that after data augmentation and feature extraction, the XGBoost model we proposed outperform the LightBGM and the traditional GBRT and Random Forest.

KEYWORDS

percipitation prediction, XGBoost, radar echo map

ACM Reference Format:
Anqi Liu. 1997. Combined XGBoost Model for Short-term Percipitation Nowcasting. In *Proceedings of ACM Woodstock conference, El Paso, Texas USA, July 1997 (WOODSTOCK'97),* 3 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

Short-term precipitation prediction, especially the rainfall forecasting is a task that focus on providing an accurate prediction over a relatively short-term (several hours) rainfall intensity of a certain area. Since traditional methods for precipitation forcasting, such as numerical weather prediction (NWP), require complex and punctilious simulation of the physical equations of the atmophere model[11], and extrapolation based methods can be much more faster and more accurate, the most state-of-the-art methods used in these years are based on the radar data. Moreover, with the rapid development of Machine learning, some computer vision methods, ranging from Convolutional Neural Network (CNN) to optical flow feature extraction methods have been employed into the analysis of radar echo maps [3, 5]. These technical approches have proved the possibility of using Machine Learning models to satify the speed and accuracy requirement of short-term rainfall nowcasting.

In the CIKM AnalytiCup 2017, the data we make use of is radar echo extrapolation data which cover a central target site and its surrounding area; it is marked as a 101×101 matrix with each grid point has a radar reflectivity value and the target place located at the (50,50) point. Considering the situation of having a data set with the number of feature greater than its number of sample, we decide to use a model that more complex than Linear Regression but not complex as deep neural network. Ensemble tree model has shown its ability of dealing with time series and image data [6, 8, 9]. XGBoost is an end-to-end tree boosint model which recently has been proved to be widely mighty and highly effective in many data mining competition [1].

In this paper, we adopt a multi-XGBoost framework as by combining two XGBoost model trained from two different feature-engineered data set, and utilize this difference to fuse the analysis of image and time characteristics of radar map.

2 METHODS

2.1 Data Preprocessing

The raw data was given by the .txt format covers 4 dimensions. Each radar sample contains 15 time frames, 4 height level, and each map covers a 101×101 area with the target at center. After extracting from .txt file, we transform the data into a 5-dimension (10000, 15, 4, 101, 101) *Numpy array*, we reduce the memory size by change its data type from *float32* into *int8*. As observation, the number of missing values in the radar map is very low (less than 0.01%), their effect on the prediction is negligible. So we can neglect them.

When facing the problem of having less samples but much more features, the data augmentation is essential. We augment the data set by transpose the origin matrix, rotate and flip both of them, and then select the unique ones.

2.2 Feature Extraction

2.2.1 Height. By using matplotlib and PIL package, we draw the picture from radar maps. After observation, we found that the first height level (H_0 level) contains lots of black area which means a very low reflectivity value. We believe that this is caused by the blocking of buildings or moutains in such a low level of vertical height. Considering the other 3 level, we found that the movments and reflectivity density information in the 3_{rd} and 4_{th} levels are shown in the second level (H_1). So we decide to use only the H_1 map as the main features map.

2.2.2 Timestep. By running the XGBoost Regression model at first time, the output of feature importance shown that the last 4 time frames are the most "important" time frames and the middle

WOODSTOCK'97, July 1997, El Paso, Texas USA

ACM ISBN 123-4567-24-567/08/06...$15.00
https://doi.org/10.475/123_4

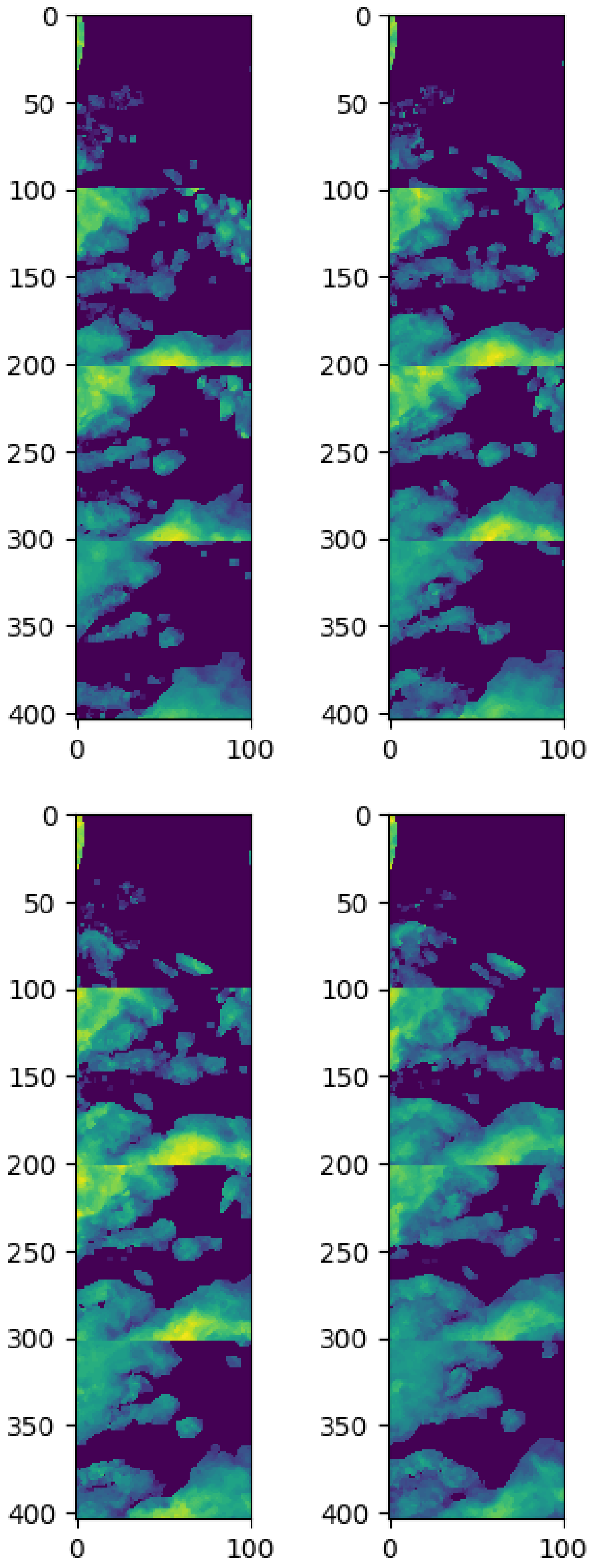

Figure 1: Samples of the picture of radar map. The four pictures are aranged by time and each contains 4 heights

Table 1: online score of the first round testing

Model	Input Feature Shape	RMSE
Random Forest	$15 \times 4 \times 101 \times 101$	14.76798174
Random Forest	$4 \times 1 \times 51 \times 51$	13.83612591
GBRT	$4 \times 1 \times 51 \times 51$	13.66662582
LightGBM	$4 \times 1 \times 51 \times 51$	13.53390808
XGBoost	$4 \times 1 \times 51 \times 51$	13.43600425
XGBoost	$4 \times 1 \times 71 \times 71$	13.46450897
XGBoost	$4 \times 1 \times 71 \times 71$(diff)	13.27905385

Table 2: online score of the second round testing

Model	Input Feature Shape	RMSE
Random Forest	$4 \times 1 \times 51 \times 51$	14.45158317
LightGBM	$4 \times 1 \times 51 \times 51$	14.16237720
XGBoost	$4 \times 1 \times 71 \times 71$(diff)	13.44212137
combined XGBoost	$4 \times 1 \times 71 \times 71$(one with diff)	13.29852754

frames are less useful. Furthermore, we think that the movement of cloud plays an important role in the rain formation. The tendency and track of movement in the following 1 hour are much more clear and can be traced and learned in the last 4 timesteps. So we reduce the data scaler by using only the last four time frames.

2.2.3 Difference. Since XGBoost only accepts the sample as a feature vector, the input data set should be reshaped into a 2 dimensions matrix. When it comes to the time series, simple vectors cannot provide the time dimension well. Then differencial features are necessary. We use the method of taking difference between two adjacent time windows, the later one minus the former one, as a new feature matrix. The last 4 time frames create 3 difference matrices. Together with the last one (15_{th}) timesteps are a new traing set for the second XGBoost model.

2.3 XGBoost model

XGBoost is a scalable end-to-end tree boosting model proposed by Tianqi Chen *et al.* in 2015. After publishment, XGBoost quickly take up an crutical position in the data competition. Lots of the winners employed this model when participating the Kaggle data challenge. It has been used in high-dimensional databases, emotion recognition in music task, language network and so on [2, 4, 7, 10]. Taking the advantage of highly effective and being able to handle high-dimensional data set, XGBoost outperfoms the traditional models such as Gradient Boosting Regression Tree in scikit-learn, Random Forest, and even, the novel model proposed by Microsoft - LightGBM.

3 EXPERIMENT AND COMPARISON

We first process the data as said in section 2, and then extract the central 51×51 and 71×71 area from the processed radar map as the final training set. Taking the result of Random Forest as baseline in first round. As the result shown in Table 1, we found that XGBoost models with 51×51 or 71×71 input shape exceed the other models. And the differential feature can greatly impove the result.

By taking average of the output from two XGBoost models (one with 51×51 input, one with 71×71 differential input), we simply combine the models to get an regression result in second round. The results are shown in Table 2.

4 CONCLUSIONS

This paper has provided the methods and models we use during the Short-Time Quantitative Precipitation Forecasting Challenge in CIKM AnalytiCup 2017. The experiment and online scores shown demonstrate that our feature engineering methods, including data augmentation and feature extraction, along with the combined XGBoost models can significantly enhance the precipitation nowcasting accuracy in comparison with other ensemble tree methods.

Be limited by the hardware equipment, we only did some simple augmentation (The XGBoost model need to use double the memory space when training model). By rotating and fliping the origin data, the data set was augmented from 10000 to 80000. This number is still not large enough in this case. There is still much room for improvement.

As an explanation, the output between the submission on the website and the submitted code would be different. This is because of the random state and the random seed in different environment. Furthermore, we used GPU during the training and because of the memory of our GPU is not big enough, we manually mini-batch the data set. This may cause some difference in the output. However, this gap should not be large.

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[2] Tianqi Chen and Tong He. 2015. Higgs boson discovery with boosted trees. In *NIPS 2014 Workshop on High-energy Physics and Machine Learning*. 69–80.

[3] P Cheung and HY Yeung. 2012. Application of optical-flow technique to significant convection nowcast for terminal areas in Hong Kong. In *The 3rd WMO International Symposium on Nowcasting and Very Short-Range Forecasting (WSN12)*. 6–10.

[4] Yuchao Fan and Mingxing Xu. 2014. MediaEval 2014: THU-HCSIL Approach to Emotion in Music Task using Multi-level Regression.. In *MediaEval*.

[5] Benjamin Klein, Lior Wolf, and Yehuda Afek. 2015. A dynamic convolutional layer for short range weather prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4840–4848.

[6] Robert K Lai, Chin-Yuan Fan, Wei-Hsiu Huang, and Pei-Chann Chang. 2009. Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications* 36, 2 (2009), 3761–3773.

[7] Francisco Javier Martinez-de Pison, Esteban Fraile-Garcia, Javier Ferreiro-Cabello, Rubén Gonzalez, et al. 2016. Searching parsimonious solutions with GA-PARSIMONY and XGBoost in high-dimensional databases. In *International Conference on EUropean Transnational Education*. Springer, 201–210.

[8] Christopher Meek, David Maxwell Chickering, and David Heckerman. 2002. Autoregressive tree models for time-series analysis. In *Proceedings of the 2002 SIAM International Conference on Data Mining*. SIAM, 229–244.

[9] Ping Tan, Gang Zeng, Jingdong Wang, Sing Bing Kang, and Long Quan. 2007. Image-based tree modeling. In *ACM Transactions on Graphics (TOG)*, Vol. 26. ACM, 87.

[10] L Torlay, M Perrone-Bertolotti, E Thomas, and M Baciu. 2017. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Informatics* (2017), 1–11.

[11] Shi Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*. 802–810.