ICIC Express Letters
Volume **10**, Number **1**, January **2016**
 ISSN 1881-803X
pp. **137–144**

A CONDITIONAL RANDOM FIELD MODEL FOR REAL-TIME URBAN AIR QUALITY FORECAST

Yuan Huang, Guoyan Huang and Jiadong Ren

College of Information Science and Engineering
Yanshan University
No. 438, Hebei Avenue, Qinhuangdao 066004, P. R. China
757918272@qq.com

Received July 2015; accepted October 2015

Abstract. *In view of the nonlinearity of air quality forecast and the influence of meteorological factors, a conditional random field (CRF) model for real-time urban air quality forecast is proposed in this paper. Some meteorological features are extracted on the basis of analysis on the influence of meteorological factors on air quality. The conditional random field model is employed to forecast real-time urban air quality index (AQI) levels with extracted meteorological feature data, the AQI levels of corresponding time and real-time meteorological forecast data. A novel feature template and Model updating rule are defined during the forecast of CRF model in order to respectively improve forecast accuracy and guarantee efficiency. Experimental results show that with multiple meteorological features taken into consideration, the CRF model is suitable for forecasting AQI levels due to less error and higher accuracy.*
Keywords: Air quality, Meteorological factors, CRF, Forecast

1. **Introduction.** Information about urban air quality is of great importance to protect human health and governments policy making. Air quality index (AQI) is a number used by government agencies to communicate to the public how polluted the air is currently. AQI is divided into six levels according to the size of the value, which include G, M, U-S, U, VU and H, as shown in Figure 1. The greater the AQI values are, the darker the color of the corresponding level is, showing the air pollution is more serious. The AQI levels make people understand the air quality information in a clearer and simpler way, and help people better protect health. Valuable reference information can be provided for the public and government decision-making in time by forecasting AQI levels in real time.

The bulk of existing work on the statistical forecasting of air quality is based on either neural networks or grey models [1, 2]. Wang et al. [3] utilized a hybrid artificial neural network to enhance the forecast accuracy by revising the error term of the traditional method. Zheng et al. [4] combined artificial neural network with linear regression to infer the air quality. Pai et al. [5] utilized GM$(1, 1)$ model of grey models to forecast hourly PM_{10} and $PM_{2.5}$ concentrations in Banciao city of Taiwan. However, neural networks and grey models are both subject to important drawbacks. The neural networks are complicated, prone to in-sample over fitting, and easy to fall into local minimization problem, with low convergence speed and poor real-time performance. Grey models are deficient in the abilities of self-learning and self-organization with other disadvantages that the ability of processing nonlinear big data is weak, that GM$(1, 1)$ model is of the single input and single output, and that the influence of meteorological factors on air quality is not taken into account.

The conditional random field (CRF), originally presented by Lafferty et al. [6], has the advantages of self-learning, high convergence speed, strong ability for fusing multiple features and good forecast performance for nonlinear big data and so on. CRF method is

AQI	Values Levels of Health Concern	Colors
0-50	Good (G)	Green
51-100	Moderate (M)	Yellow
101-150	Unhealthy for sensitive groups (U-S)	Orange
151-200	Unhealthy (U)	Red
201-300	Very unhealthy (VU)	Purple
301-500	Hazardous (H)	Maroon

FIGURE 1. Division of AQI level

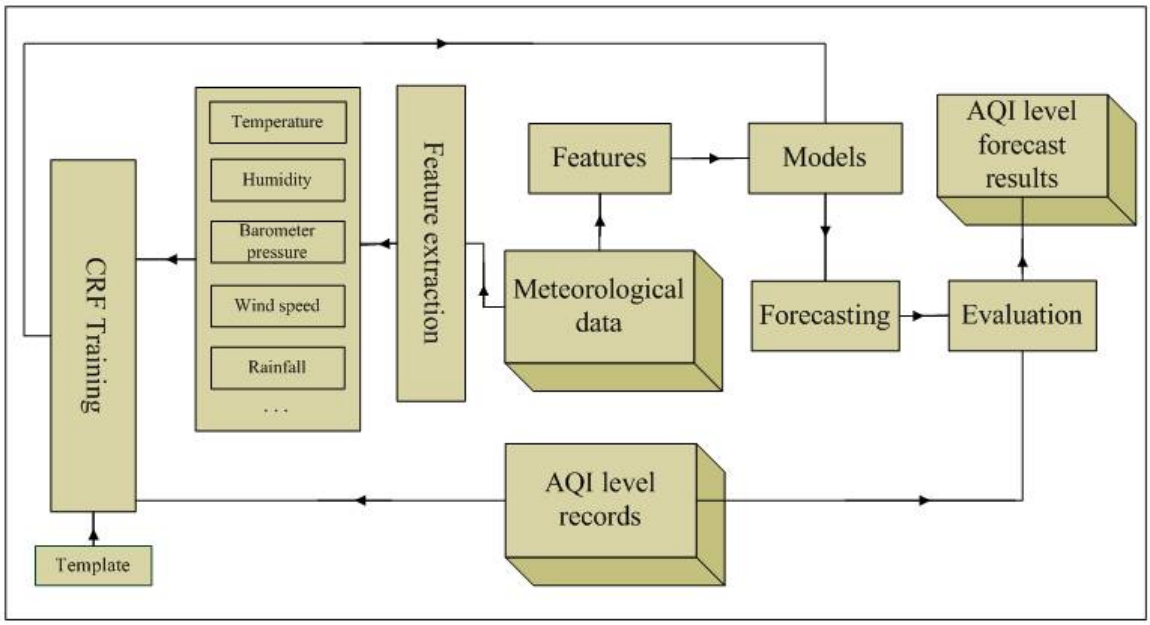

FIGURE 2. Framework of CRF model

suitable for air quality forecast due to the advantages. However, CRF method is barely used to forecast air quality in studies.

In this paper, on the basis of analyzing the influence of meteorological factors on air quality, according to the urban meteorological data and the AQI levels of the corresponding time, the CRF method is employed to forecast real-time air quality. In order to improve the forecast accuracy and guarantee the forecast efficiency, we respectively define a novel feature template (FT) and Model updating rule (MUR) in the forecast process of CRF model. Experimental results show that the CRF method has high forecast accuracy and good real-time performance.

The remaining of the paper is organized as follows. Section 2 gives the framework of the CRF model. The meteorological features are extracted in Section 3. The establishment of the CRF model is introduced and the feature template and Model updating rules are defined in Section 4. Section 5 is experiments. The paper is concluded in Section 6.

2. **The Framework of the CRF Model.** The framework of the CRF model is shown in Figure 2. Firstly, urban meteorological data are analyzed and some meteorological features are extracted. Secondly, according to Figure 1, AQI values are converted to corresponding AQI levels. The extracted meteorological data and the AQI level data of the corresponding time are used as the training data set of CRF model. Thirdly, a novel feature template is defined based on the training data set, and the features are extracted from the training set according to the feature template. Fourthly, the CRF

learning algorithm is used to learn training data set and generate the *Model*. The forecast accuracy is compared between the newly generated Model and those at earlier time points, as a result of which the *Model* with the highest accuracy is selected. Finally, the real-time meteorological features are selected as a test set. The AQI levels of corresponding time are forecasted by CRF model. The forecast levels are compared with the real ones, according to which the forecast accuracy and recall are calculated.

3. **Meteorological Feature Extraction.** Urban air quality is directly affected by the local air pollutant emissions and meteorological factors. The meteorological factors are the main factors under the condition that pollutant emissions are relatively stable [7]. The meteorological features are taken into account in the forecast, which will effectively improve the accuracy of the air quality forecast. According to China's air quality report and monitoring analysis of municipal air monitoring stations, the meteorological factors affecting the air quality mainly include temperature, wind speed , atmospheric pressure and so on. Accordingly, we identify five features: atmospheric pressure (F_P), temperature (F_T), rainfall (F_R), wind speed (F_W) and humidity (F_H) on the basis of considering factors like the convenience of getting data and the importance of impact on air quality. Table 1 is the monthly correlation coefficients between AQI values and meteorological factors in Beijing, Tianjin and Shijiazhuang in the past ten years. Table 1 indicates there is the significant positive correlation between the atmospheric pressure and AQI values and the negative correlation between other meteorological features and AQI values. In Table 1, P denotes Pearson coefficient, and M denotes the number of months.

Table 1. Monthly correlation coefficients between AQI and meteorological factors

		Atmospheric pressure (hPa)	Temperature (°C)	Rainfall (mm)	Humidity (%)	Wind speed (m/s)
Beijing	P	0.183	−0.266	−0.431	−0.359	−0.383
	M	120	120	120	120	120
Tianjin	P	0.383	−0.470	−0.372	−0.258	−0.050
	M	120	120	120	120	120
Shijiazhuang	P	0.404	−0.471	−0.407	−0.087	−0.075
	M	120	120	120	120	120

Atmospheric pressure and AQI The air pollutants rise to high altitude when atmospheric pressure is low. The lower the pressure is, the better it is for air pollutants to diffuse and dilute, with corresponding AQI decreased.

Temperature and AQI The higher the temperature is, the more intense air convection activity is, with the results that it is more conducive to the diffusion of air pollutants and that air quality is better.

Rainfall and AQI Rainfall can effectively remove and wash air pollutants, and reduce the concentrations of various air pollutants, which will purify the air.

Wind speed and AQI The greater wind speed is, the more conducive it is to the dilution and diffusion of pollutants, with better air quality.

Humidity and AQI The high humidity indicates the emergence of rain and snow, which can effectively reduce the concentrations of air pollutants.

4. The Air Quality Forecast Model.

4.1. CRF model.

In considering the influence of the meteorological features on air quality, we use linear-chain CRF to forecast the real-time AQI levels. The advantage of CRF over hidden Markov models is the relaxation of the independence assumptions between

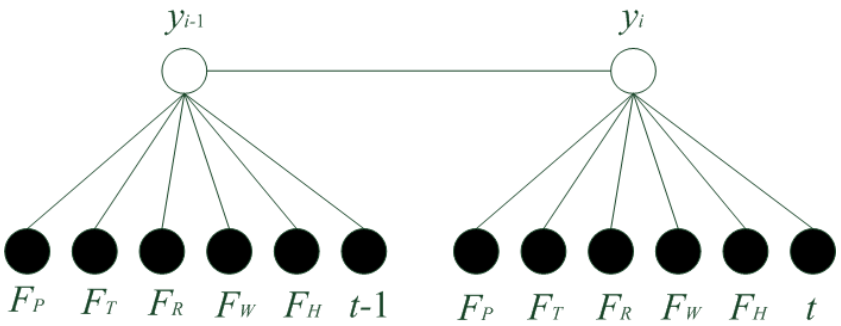

FIGURE 3. Structure of CRF forecast model

features. Additionally, CRF avoids the label bias problem exhibited by maximum entropy Markov models. CRF defines the conditional distribution of a label sequence Y when an observation sequence X is given.

As shown in Figure 3, the white nodes $Y = y_1, y_2, \ldots, y_m$ represent hidden state variables to be inferred given the sequence of observations denoted by black nodes $X = x_1, x_2, \ldots, x_m$, where $x_i = F_P, F_T, F_R, F_W, F_H, t$ (t is a timestamp by hour, e.g., 9am). The y_i is structured to form a chain with an edge between each y_{i-1} and y_i, as well as has an AQI label belonging to $G, M, U\text{-}S, U, VU, H$. When conditioned on X, the random variables y_i obey the Markov property with respect to the graph: $P(Y_i|X, Y_j, i \neq j) = P(Y_i|X, Y_j, i \sim j)$, where $i \sim j$ means that i and j are neighbors in graph. (X, Y) is a conditional random field.

The probability of a particular label sequence y given observation sequence x is defined as a normalized product of potential functions as follows:

$$p(y \mid x) \propto \exp\left(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i)\right) + \sum_k \mu_k s_k(y_i, x, i) \tag{1}$$

where $t_j(y_{i-1}, y_i, x, i)$ is a transition feature function of the entire observation sequence and the label at positions $i-1$ and i; $s_k(y_i, x, i)$ is a state feature function of the label at position i and the observation sequence; λ_j and μ_k are parameters to be estimated from training data.

The two feature functions are unified as: $f_i(y_{i-1}, y_i, x, i)$, and we transfer Equation (1) to [6]:

$$p(y \mid x) \propto \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i)\right) \tag{2}$$

where $Z(x) = \exp\left(\sum_j \lambda_j t_j\left(y_{i-1}, y_i, x, i\right)\right)$.

The CRF algorithm framework is presented in Algorithm 1.

4.2. **The novel feature template.** CRF model can effectively deal with the relationship between the multiple features through feature template without considering independence assumption. Different feature templates will affect the forecast accuracy of the CRF model. According to the relationship between the meteorological features and the air quality, we define a novel feature template, as shown in Figure 4, to achieve high feature recognition effect and high forecast accuracy of the algorithm.

The five meteorological features are defined as uncorrelated, namely five features do not affect each other, and the fact is taken into account that meteorological features at a previous time point exert no effect on those at the next time point. Real-time AQI levels are inferred only by five meteorological features at the current time point. In Figure 4, each line is a template. Each template is specified by $\%\mathbf{x}[\mathbf{row}, \mathbf{col}]$ to specify a token in

Algorithm 1: CRF

Input: A set of meteorological features (F_p, F_T, F_R, F_W, F_H) and a set of AQI values
Output: Forecast AQI levels
1. Do
2. AQI values → AQI levles
3. CRF_learn (F_p, F_T, F_R, F_W, F_H, AQI levels)
4. The feature template (FT) is defined
5. Apply FT to extracting features from training data set
6. Generate Model and compare the forecast accuracy of Models
7. Select the Model with the highest accuracy
8. CRF_test (F_p, F_T, F_R, F_W, F_H)
9. Return Forecast AQI levels

```
U00:%x[0,1]
U01:%x[0,2]
U02:%x[0,3]
U03:%x[0,4]
U04:%x[0,5]
U05:%x[0,6]
```

FIGURE 4. The novel feature template

TABLE 2. *Models* forecast accuracy comparison

	Forecast accuracy (9am)	Forecast accuracy (10am)
Model-1	83%	87%
Model-2	82%	83%

the input data. The row specifies the **row** offset of the current token, and the **col** the column location. $\%\mathbf{x}[\mathbf{0},\mathbf{1}]$ represents atmospheric pressure feature, $\%\mathbf{x}[\mathbf{0},\mathbf{2}]$ represents temperature feature, $\%\mathbf{x}[\mathbf{0},\mathbf{3}]$ represents rainfall feature, $\%\mathbf{x}[\mathbf{0},\mathbf{4}]$ represents wind speed feature, $\%\mathbf{x}[\mathbf{0},\mathbf{5}]$ represents humidity feature, and $\%\mathbf{x}[\mathbf{0},\mathbf{6}]$ represents the AQI level. The five meteorological features are extracted through $\%\mathbf{x}[\mathbf{row},\mathbf{col}]$ in a large number of meteorological data.

4.3. **The model updating rule.** CRF model will generate a *Model* in every forecast through the training data set and the feature template. *Model* is directly related to the forecast performance of CRF, and the appropriate *Model* will help to improve the forecast accuracy of CRF.

Table 2 is a comparison of forecast accuracy between the *Models* at 9am and 10am in Beijing on June 20, 2015. *Model*-1 is generated at 9am and *Model*-2 at 10am. Table 3 presents the generated *Model* at the current time point, which is not the one with the highest accuracy and the generated Model at an earlier time point may have higher forecast accuracy.

Therefore, we select the *Model* with the highest accuracy by comparing the accuracy between the generated *Model* at current time point and the ones at the earlier time points in the real-time forecast. However, every real-time forecast will produce a *Model*. The accumulation of numerous *Models* will reduce the efficiency of the CRF forecast model which requires regularly updating the accumulated *Models*. For *Model* update setting, excessively quick update will lead to the deletion of *Model* with high forecast accuracy. Excessively slow update will affect the efficiency of CRF model. Accordingly, we define the Model updating rule (*MUR*) through contrast experiment in order to simultaneously guarantee the forecast accuracy and efficiency of CRF model.

Definition 4.1. *Model updating rule* (MUR) *is defined as Equation (3):*

$$MUR = 30(N-1) \tag{3}$$

where N *is the number of forecast during the day. Namely, update is carried out once every 30 days, which eliminates the earliest* $30(N-1)$ *Models.*

The design of the feature template and *Model updating rule* helps improve the forecast accuracy of AQI level and guarantee the efficiency of the CRF model.

5. Experiments.

5.1. **Datasets.** In the evaluation, we take Beijing as an example to forecast the AQI level. The hourly meteorological data, consisting of atmospheric pressure, temperature, rainfall, wind speed and humidity, is collected from the public meteorological website records during a year. The AQI data of corresponding time is from the air quality real-time release system of Beijing municipal environmental monitoring center. We use the real datasets detailed in Table 3. As the air quality real-time release system may not have reports sometimes, we present the hours of effective records in Table 3.

Table 3. Details of the datasets

Meteorological data	Hours Time Span	8650 5\31\2014 5\31\2015
AQI data	Hours Time Span	8650 5\31\2014 5\31\2015

5.2. **Results and analysis.** We use a half of the data for training and the rest for testing, ensuring both parts of data have a relatively balanced distribution over different AQI levels. After the forecast is completed, the training set and the testing set are exchanged to carry out another forecast. The forecast results of two forecasts are combined. The forecast accuracy and the recall are calculated as shown in Table 4.

Table 4 presents the mean forecast accuracy as high as 77.6% which is obtained by using our CRF model to forecast the AQI levels of Beijing. The forecast accuracy of G level is the highest, up to 88.5%. Although the forecast accuracy of $VU\&H$ level is the lowest, it reaches 70.4%. The mean recall of CRF model is 81.6%. Through the analysis, it can be concluded that our CRF model is suitable to forecast the future urban air quality, with good forecast performance. Our CRF model is employed to forecast the AQI levels of Beijing from 0am to 23pm on June 20, 2015. The forecast results are shown in Figure 5.

In Figure 5, only three time points in the AQI level forecast do not match with the real AQI levels, respectively at 7am, 9am and 18pm. At 7am and 9am, the AQI level (M) is lower than the real AQI level (U-S), and the AQI level (U-S) is higher than the real AQI level (M) at 18pm. The forecast accuracy of CRF method is 87.5%, which further indicates that the CRF method has high forecast accuracy under the consideration of the influence of multiple meteorological features.

We use the artificial neural network (ANN) method to forecast the AQI levels of Beijing with the same datasets, and compare the forecast accuracy and the recall with those of our CRF method. Table 5 shows the comparison results of forecast accuracy and recall. Results show that in terms of forecast accuracy and recall, the CRF method is superior to the neural network method.

The efficiency of our approach, which was tested on a 64-bit server with an AMD A8-3850 2.90 GHZ CPU and 16GB RAM. On average, we can complete one forecast in 12.6ms. One forecast is completed in 13.1ms by the traditional CRF model. With good real-time performance, our CRF model is suitable for the urban air quality forecast.

TABLE 4. The forecast accuracy and recall of CRF

Real level	Forecast level					Recall
	G	*M*	*U-S*	*U*	*VU&H*	
G	1886	423	9	0	0	0.814
M	243	1882	342	33	0	0.753
U-S	3	270	1190	208	69	0.684
U	0	0	123	1364	129	0.844
VU&H	0	0	0	6	470	0.987
Accuracy	0.885	0.731	0.715	0.847	0.704	

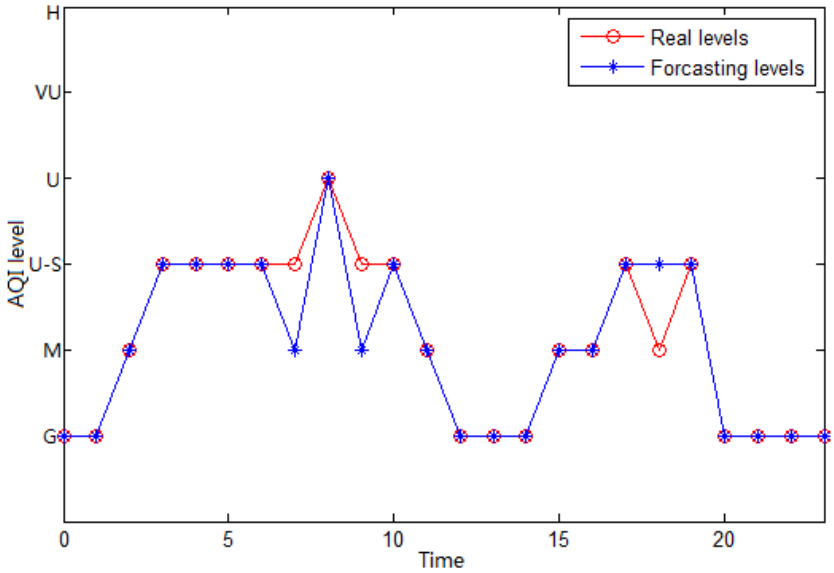

FIGURE 5. AQI levels forecast

TABLE 5. Comparison results of forecast accuracy and recall

	ANN	CRF
Accuracy	77.6%	64.5%
Recall	70.4%	56.4%

6. **Conclusions.** In this paper, the CRF model is applied to forecasting the AQI levels based on the consideration of the influence of multiple meteorological features on air quality. We respectively define the feature template and Model updating rule to extract meteorological features and guarantee the efficiency of the CRF model, and take Beijing as an example for experiments and analysis. Experimental results show that mean forecast accuracy and recall of CRF are high in forecasting the AQI levels. Comparison with the ANN method indicates the forecast accuracy and the recall of our CRF method are respectively 13.1% and 15.0% higher than those of the ANN method. These results demonstrate with high forecast accuracy, and our approach is comprehensive and simple. In the future, we would like to apply the CRF model to more cities, further improve the forecast accuracy and study the root causes of air pollution.

Acknowledgment. This work is supported by the National Natural Science Foundation of China under Grant No. 61170190 and No. 61472341, and the Natural Science Foundation of Hebei Province P. R. China under Grant No. F2012203062, No. F2013203324

and No. F2014203152. Authors also gratefully acknowledge the helpful comments and suggestions of reviewers, which have improved the presentation.

REFERENCES

[1] J. Westerlund, J.-P. Urbain and J. Bonilla, Application of air quality combination forecasting to Bogota, *Atmospheric Environment*, vol.89, pp.22-28, 2014.

[2] L. Pan, B. Sun and W. Wang, City air quality forecasting and impact factors analysis based on grey model, *Proc. of the 2011 SREE Conference on Engineering Modeling and Simulation*, Hong Kong, China, 2011.

[3] P. Wang, Y. Liu, Z. Qin and G. Zhang, A novel hybrid forecasting model for PM_{10} and SO_2 daily concentrations, *Science of the Total Environment*, vol.505, pp.1202-1212, 2015.

[4] Y. Zheng, X. Yi, M. Li et al., Forecasting fine-grained air quality based on big data, *Proc. of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, Sydney, Australia, 2015.

[5] T.-Y. Pai, C.-L. Ho et al., Using seven types of GM$(1,1)$ model to forecast hourly particulate matter concentration in Banciao City of Taiwan, *Water, Air, and Soil Pollution*, vol.217, pp.25-33, 2011.

[6] J. Lafferty, A. McCallum and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proc. of the 18th International Conference on Machine Learning*, Williamstown, USA, pp.282-289, 2001.

[7] Z. Zhou, S. Zhang et al., The impact of meteorological factors on air quality in the Beijing-Tianjin-Hebei region and trend analysis, *Resources Science*, vol.36, pp.191-199, 2014.