Prediction of nitrogen oxides Emission from Coal - fired Boiler Based on RF - GBDT

Liangming Gui [1a] Yongjun Xia [1], Haishan Li [1], Peng Tan [2],

Shangzhi Zhang[2b], Cheng Zhang [2]

1. State Grid Jiangxi Electric Power Research Institute, Jiangxi Province, 330096

2. State Key Laboratory of Coal Combustion (Huazhong University of Science and Technology), Wuhan 430074, China

[a]email:23507104@qq.com, [b]email: 502631161@qq.com

Key words: nitrogen oxides prediction; random forest (RF); gradient lift decision tree (GBDT); coal fired boiler

Abstract: A nitrogen oxides emission forecasting model was established for a supercritical 660MW unit boiler combined with random forest (RF) and gradient lift decision tree (GBDT) algorithm. The steady-state working point of the historical data is screened from the SIS system of the power plant. The feature feature of the RF model is used to filter the data characteristics, and the GBDT model for predicting nitrogen oxides emission is established with the selected feature as the input variable. The comparison with support vector machine (SVM), RF and other models shows that RF-based feature selection can improve model performance. Compared with other models, RF-GBDT has the highest prediction accuracy of nitrogen oxides emission.

Introduction

Coal-fired power plants are an important component of China's energy system. The nitrogen oxides (NOx) produced by coal combustion endanger the human body and the environment, causing damage to the ecosystem, which is one of the pollutants that the thermal power plant needs to control [1].

Low NOx combustion optimization from the root causes of the problem of NOx emissions, compared with the follow-up treatment, with easy operation, strong and so on [2]. On the other hand, with the changes brought about by the data age, information mining and utilization are also integrated into the work of the boiler boiler combustion optimization. As a product of both, low NOx combustion optimization based on historical data of power plant has become widely studied in the field. This method uses the idea of data mining to establish the NOx emission model by using the real-time operation record of the power plant unit. The prediction results and the optimization of the operating parameters are used to guide the actual operation of the operation personnel. The data source is provided by the power station SIS system. Compared with the traditional modeling and combustion optimization based on the thermal combustion adjustment test data, the method of relying on historical data modeling has the characteristics of small workload, comprehensive data and high resource utilization.

The prediction model has undergone the development of "decision tree - neural network - support vector machine", which does not represent the merits of the model. By rational use and optimization, the classic model may have better performance. The "NOX emission prediction model based on RF and GBDT" (RF-GBDT) proposed in this paper is the optimization product of classical decision tree model under the model joint idea. In this paper, a large number of historical operation records are obtained from the SIS system, and the steady state condition data is obtained by preprocessing. The feature of the RF learning model is selected to select the data characteristics. After the feature screening is finished, the GBDT algorithm is used to establish the prediction model of NOx emission in the boiler. In order to evaluate the performance of the model, the RF-GBDT model is compared with the effect of SVM and RF model.

model introduction

Random Forest (RF)

Random forest originated from the concept of Random Decision Forests proposed by Tin Kam Ho in 1995. In 2001, Breiman first proposed random forests and related algorithms [9].

The random forest model is a typical representative of a strong classifier composed of multiple weak classifiers. The model combines a large number of classification regression trees (CART) with Bagging ideas to achieve the effect of enhancing model performance. The flow of ideas as shown in Figure 1 [9]: by re-sampling method to extract multiple samples, each sample is used to establish a "weak" decision-making model, and finally set the decision tree decision, through the final decision Of the forecast results. There are related literatures that, on the classification problem, the random forest algorithm has higher precision than the Bagging classification tree [10].

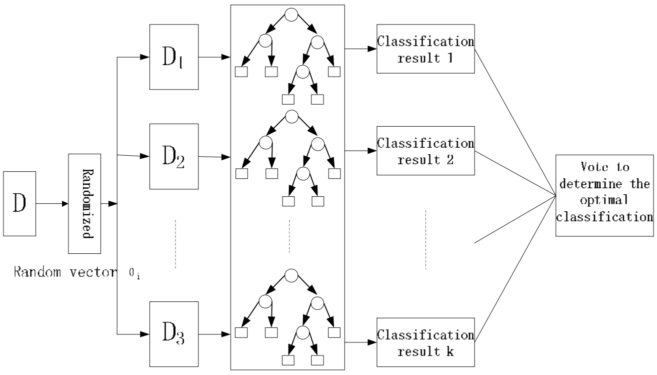

Fig. 1 Method of the RF model

gradient decision tree (GBDT)

The gradient lift decision tree (GBDT) was first proposed by Friedman [11] in 2001, and the subsequent improvements and improvements are based on the core idea of gradient boosting. GBDT, together with SVM, is considered to be an algorithm with strong generalization ability, and has been attracting attention in recent years because of the machine learning model used in search sorting [12,13].

GBDT Essence is a combination of a large number of simple models. The GBDT used in modeling and the RF in the above are the combined model of decision tree algorithm, which is different in the way of model combination. RF model is a large number of decision trees to produce independent, parallel decision-making, and then aggregated to form the final result; and GBDT model is the decision tree unit "series": decision tree arrangement with strict timing sequence, each decision tree for The input information of the modeling is the sum of the output information of the upstream tree (except for the first tree). Through this boosting method based on boosting ideas, a large number of decision-making units together to get the final output of the model results.

The pseudo-code of the idea is as follows [11]:

```
Algorithm : Gradient_Boost
$F_0(x) = \arg\min_r \sum_{i=1}^{N} L(y_i, r)$
$For\ m = 1$ to $M$ do:
    $\tilde{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, N$
    $a_m = \arg\min_{a,b} \sum_{i=1}^{N} [\tilde{y}_i - b\,h(x_i : a)]^2$
    $r_m = \arg\min_r \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + r\,h(x_i : a_m))$
    $F_m(x) = F_{m-1}(x) + r_m h(x : a_m)$
$endFor$
$end$ Algorithm
```

NOx emissions modeling

This model uses more than 50,000 data from the model, from a 660MW ultra-supercritical power plant boiler. Data attribute characteristics of more than 100, respectively, corresponding to the SIS system in the unit of the throttle opening, swing angle, coal mill instant parameters, as well as a number of indicators parameters. Data preprocessing for the above data is divided into the following steps:

1) to deal with missing points caused by abnormal work points, abnormal values;

2) eliminate the abnormal operation of the boiler data and fuel combustion combustion conditions data;

3) Based on the parameters such as boiler load and total wind, the steady-state condition data for modeling are screened out.

Fast selection of 5000 data from the preprocessed data set to form the modeling data set. According to the idea of cyclic cross validation, the modeling data set is randomly divided into 5 equal parts. Four of the data are composed of training subsets, which are used to train the model and the fifth data is used as the test subset to validate the model. The test set is unknown to the model established using the training set and can reflect the merits of the model more reliably. Cross-validation effectively reduces the instability of the predictive set itself as a result of the instability of the model results.

In the modeling process, too much input will cost analysis and model training to spend more time; the other hand, it may lead to "dimension disaster", so that the model is overly complex and reduce its generalization ability [14]. In this paper, we use RF model feature sorting feature to select feature.

The training set and the test set are normalized, the NOx emission value is used as the model output, and the other operating parameters are input as model, and the RF model is established. The number of tree node preset variables is set to the default value (the square root of the input dimension), and the number of trees is set to 500.

The results shown in Fig. 2 are obtained by measuring the importance of the input parameters for NOx emissions in the RF model from three different aspects of average precision drop, average mean square error drop, and standard error importance. A total of 19 modeling features were selected, including: primary air, secondary air, SOFA wind and other damper opening adjustment, as well as boiler load, boiler total air volume, total fuel flow and other parameters. The results of the RF characterization show that the importance of the six mill parameters is weak. The reason is because the boiler in order to ensure good operation, the various links between the adjustment, the relevant parameters of the coal mill often determine the parameters such as the throttle opening, swing parameters and other parameters of the adjustment. Therefore, the influence of the relevant parameters of the pulverizer on the NOx emissions of the boiler is characterized more explicitly by other parameters associated with it.

The selected feature is used as input and the GBDT model is established with the NOx emission concentration as the output. In the selection of model parameters, the genetic algorithm (GA) is used to optimize the GBDT modeling parameters. The GBDT model is the union of CART, the depth and the number of trees are integers, the two genetic algorithm optimization of the work domain needs to be limited to the integer range. Based on the above series of work, the final establishment of power plant boiler NOX emission prediction RF-GBDT model.

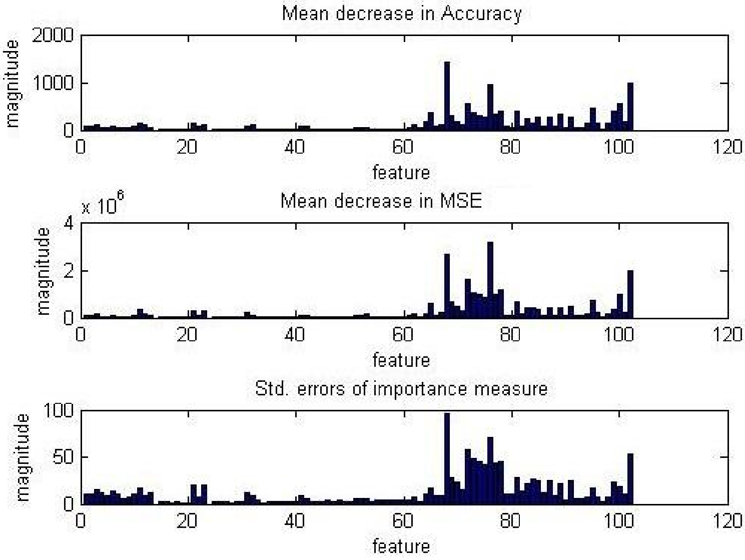

Fig. 2 Parameters for NOx emissions

Results and discussion

RF-GBDT model prediction effect

The comparison between the predicted results of the RF-GBDT model and the actual values of the test set is shown in Fig. Figure 4 is the error distribution of the model predictions. It can be seen from these two graphs that the prediction results coincide with the measured results and the prediction error is small. The average error of the model to the test set is 1.84%, and the prediction error of more than 94% is within 5%. The prediction results show that the RF-GBDT model has good prediction accuracy for NOX emission prediction modeling and strong model generalization ability.

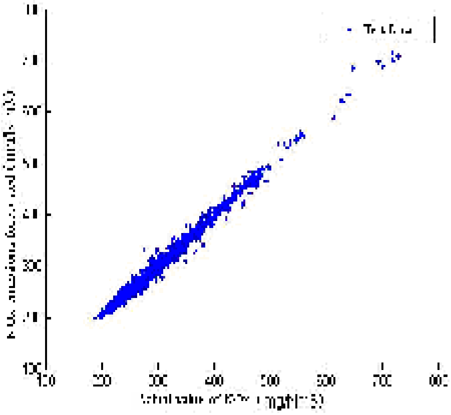

Fig. 3 Prediction result of the RF-GBDT model

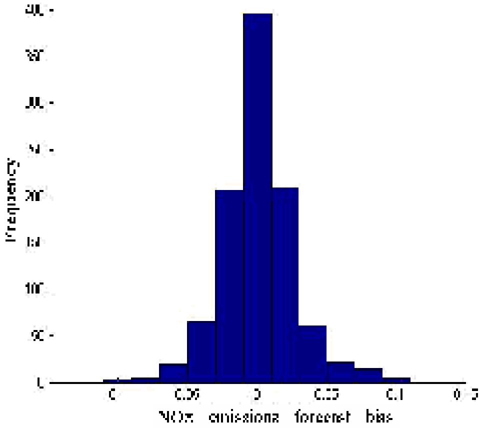

Fig. 4 Fractional Error of the RF-GBDT model

Comparison of RF-GBDT models with SVM and RF

In order to compare the RF-GBDT model with other modeling methods, other models are established and predicted using the data in the above work. Comparison of the use of a total of five models, as shown in Table 2.

Table 1 List of models

	GBDT algorithm	RF algorithm	SVR algorithm
Based on RF feature selection	RF-GBDT	2#RF	RF-SVR
Not based on RF feature selection	GBDT	RF	SVR

The modeling of the contrast model is similar to that of the RF-GBDT, using the same cross validation data packet to avoid the differences in training / test sets. In the work of modeling parameters optimization, GBDT and RF-GBDT model are consistent, using the local integer domain GA genetic algorithm to find the optimal parameters; SVR and RF-SVR using ordinary GA algorithm optimization; 2 # RF and RF Mode parameters for the two, using the grid method traversal optimization. Figure 5 is the RF feature selection data training out of the RF and SVM model predictive effect, compared with Figure 3.

Figure 6 shows the relative error distributions of the predictions of GBDT, RF and SVR models after RF feature selection. All three models have good prediction effect. In comparison, the RF-GBDT model is 41.3%, 66.1% and 94.1% in the range of ± 1%, ± 2% and ± 5%, respectively, which is higher than the 2 # RF model (37.8%, 64.9 %, 93.8%) and the RF-SVR model (40.0%, 65.2%, 93.8%) at the same accuracy.

Table 3 summarizes the performance indicators for each model. It can be seen from the table that the model based on RF feature selection is better than the model with no feature selection in the prediction result correlation, relative error, square error and time consuming. This result confirms the necessity of feature selection. RF-GBDT is superior to other models in the accuracy index, indicating that in the current data set, RF-GBDT model can achieve higher prediction accuracy. In short, the GBDT model is time-consuming for other models, although RF-based feature selection can be time-consuming, but overall, the time required for GBDT and RF-GBDT modeling is much longer than that of SVR and RF model.

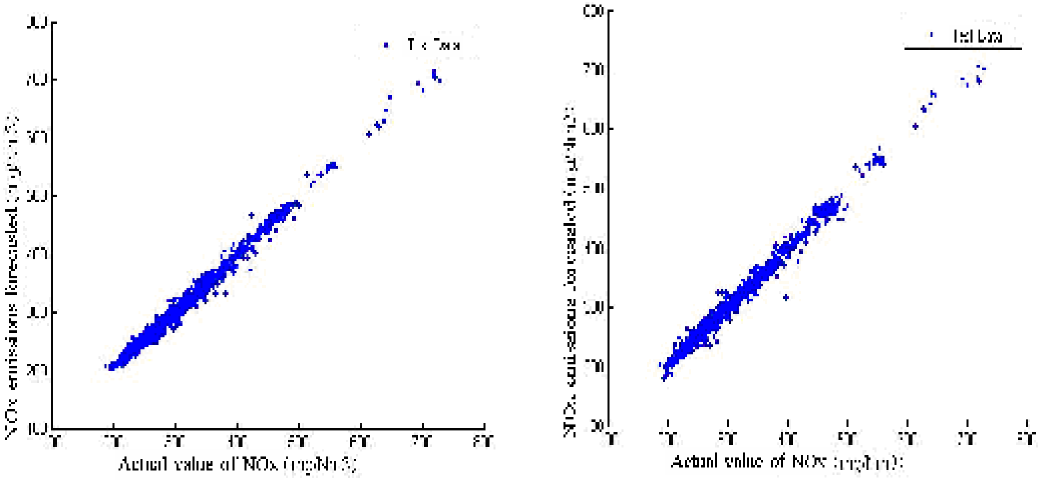

Fig. 5 Prediction result of the 2#RF model (left) and RF-SVR model (right)

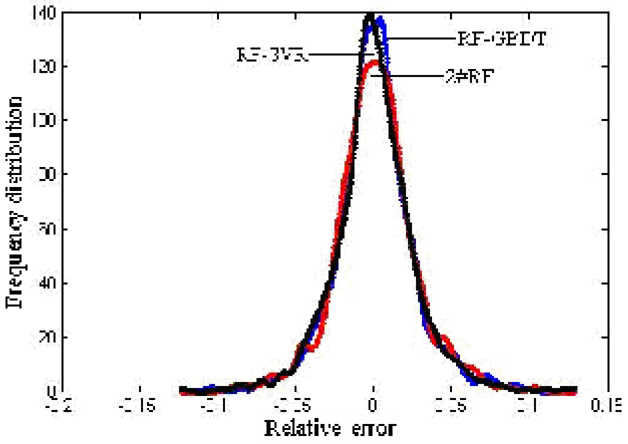

Fig. 6 Comparison of Fractional Error Distribution

SVR model takes the shortest time in the three, the accuracy is relatively poor. RF models are in the middle of both precision and time-consuming.

Table 2 Summary of various models for NO$_X$ emissions modeling

	Correlation coefficient Scc	Average relative error MRE	Average squared error MSE	Modeling time T(s)
SVR	0.986	2.12%	90.28	100.3
RF	0.987	2.22%	84.19	242.3
GBDT	0.990	1.86%	62.26	1809
RF-SVR	0.989	1.91%	71.70	43.7
2#RF	0.991	1.91%	62.47	116.4
RF-GBDT	0.991	1.84%	59.50	1286

Conclusion

In this paper, a new model of decision tree combining model: GBDT and RF is introduced for the prediction of NOx emission from a coal-fired power plant unit. A gradient-based decision tree model based on random forest feature selection is proposed by means of model fusion. -GBDT. Compared with other reference models, the model has the highest prediction accuracy and good generalization performance. The longitudinal comparison kernel algorithm, GBDT model and RF model can obtain better prediction accuracy than SVM model. However, compared with other models, RF-GBDT model calculation takes a long time. Therefore, applying this model to the on-line NOx prediction relative to the SVM and RF models requires better hardware support; or a reasonable reduction in accuracy to achieve a balance between accuracy and time.

references

[1] Muzio L J, Quartucy G C. Implementing NOX control: Research to application[J]. Progress in Energy and Combustion Science, 1997, 23(3):233-266.

[2] Li Ming, Ou Zongxian. Prediction of NOx Emission Concentration in Tangentially Fired Boiler [J]. Thermal Power Generation, 2012, 41 (2): 12-15.

[3] FANG Kuang-nan, WU Jian-bin, ZHU Jian-ping, et al.Study on the study of random forest method [J] .Journal of Statistics and Information, 2011, 26 (3): 32-38.

[4] Ma Jingyi, Xie Bangchang. Comparison of random forest and Bagging classification tree for classification [J]. Journal of Statistics and Information, 2010, 25 (10): 18-22.

[5] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5):1189--1232.

[6] Bai J, Diaz F, Chang Y, et al. Cross-Market Model Adaptation with Pairwise Preference Data for Web Search Ranking.[C]// International Conference on Computational Linguistics: Posters. 2010:18-26.

[7] Zhang B, Ye G, Wang Y, et al. Finding shareable informative patterns and optimal coding matrix for multiclass boosting[J]. Proceedings, 2009, 30(2):56-63.

[8] U Jie. Review of dimensionality reduction of high dimensional data [J] .Application Research of Computers, 2008, 25 (9): 2601-2606.