A Novel Random Forest and its Application on Classification of Air Quality

Hualing Yi
School of Big Data &Software Engineering
Chongqing University
Chongqing, China
yihualing@cqu.edu.cn

Qingyu Xiong
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
xiong03@cqu.edu.cn

Qinghong Zou
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
cquzouqh@163.com

Rui Xu
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
cquxr@cqu.edu.cn

Kai Wang
School of Automation
Chongqing University
Chongqing, China
akyle@163.com

Min Gao
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
gaomin@cqu.edu.cn

***Abstract*—Air pollution has a serious impact on daily life. It is necessary to inform the air quality in time to the public in order to take measures in advance. Machine learning methods such as random forest are good at evaluating grades of air quality. We find the distribution of air data is imbalance, which leads to negative effect on random forest classifiers. We propose a random forest method based on samples grouped bootstrap to solve this problem. Then we design three sets of experiments to evaluate the performance of the proposed method. The results of experiments indicate that the proposed method presents an improvement of random forest when both apply on balance datasets. The improvement is very significant when they apply on imbalance datasets, where the new method is much better at classifying minority samples.**

Keywords—air quality, imbalance datasets, imbalance coefficient, random forest, bootstrap

I. INTRODUCTION

Recently, air pollution is becoming increasingly serious. Problems like haze occur frequently in cities. The life of urban residents is greatly affected by air quality, especially the health. Therefore, in order to reduce the impact of air pollution on health of urban residents, it is necessary to evaluate the air quality scientifically and accurately. Methods based on machine learning do well in this task due to the vast air data in history.

The traditional methods to evaluate urban air quality mainly use fixed formulas based on fuzzy mathematics to establish the mapping between air data and air quality level, which is proved to have poor performance with low tolerance and efficiency [1]. With the rapid accumulation of a large number of air data, there are some evaluation methods based on machine learning proposed by scholars. The neural network evaluation method proposed by Bai *et al*.[2] has self-learning and self-adaptive ability, but it requires amounts of data with lots of complex computation. Method mentioned in [3] is based on SVM algorithm, which is efficient on processing two-classification problems but performs badly in multi-classification tasks.

The good generalization ability obtained by random forest models has motivated random forest to be one of the most used algorithms in data mining area [4]. In this paper, we propose a method based on random forest and assess the performance by designing three sets of experiments. The results prove that this method is effective at air quality evaluation and performs better than random forest.

The rest of the paper is organized as follows. In section II, we describe the air data we used in this paper. In section III, we introduce the basic concept of random forest algorithm. In section IV, we explain the proposed algorithm. In section V, we present the experiments we design and analyze the results. Finally, section VI is devoted to the conclusions.

II. STUDIED DATASETS

A. Data Description

The air quality data used in this paper are from the website of China Environmental Monitoring Station [5]. The *dataset1* is the air data from Beijing. The *dataset2* and *dataset3* is different air data from Fangchenggang, Guangxi province. *Dataset4* is the air data from Beijing and Fangchenggang. The feature space of datasets is consisted of 6 air pollutants (*PM2.5, PM10, SO_2, NO_2 ,O_3, CO)*. Values of features are concentration values of air pollutants. Labels of datasets are grades(*1,2,3,4,5,6*) of air quality. Each grade corresponds to a level of air quality, presented in TABLE I.

B. Data Analysis

We define a variable *Ic* called imbalance coefficient to measure the imbalance degree of a dataset. The larger *Ic* is, the more unbalanced the dataset is. We describe the process to calculate the imbalance coefficient of a dataset as follows:

The Major Science & Technology Program of Guangxi (Grant No.GKAA17129002)
The Key Research Program of Chongqing Science & Technology Commission (Grant No.CSTC2017jcyjBX0025)

978-1-7281-2627-2/19/$31.00 ©2019 IEEE
DOI 10.1109/IIAI-AAI.2019.00018

- (Step1) Let $C(C>=2)$ to be the number of categories in set S. i, j is integer between 1 and $N(1<= i, j <=C)$.
- (Step2) For each category in S, count the number of samples in S which belong to this category, denoted as X_i.
- (Step3) X is the set of X_i. For every X_i, X_j in set X and $0<X_i < X_j$, the imbalance coefficient between category i and category j is defined as:

$$Ic = \frac{X_j}{X_i} \quad (1)$$

- (Step4) Let Ic_x to be the set of Ic generated in step3, the imbalance coefficient of S is defined as:

$$Ic(s) = \frac{\sum Ic_x}{C} \quad (2)$$

TABLE II shows the imbalance degree of 4 datasets used in this paper. *Dataset2* is the most unbalanced with the largest imbalance coefficient among 4 datasets. *Dataset3* is the second unbalanced, followed by *dataset1*. While the *dataset4* is the most balanced among 4 datasets.

III. THE ORIGINAL RANDOM FOREST ALGORITHM

A. Basic Theory of Random Forest

Random Forest (RF) [6] is a fine supervised classification method based on the combination of the Breiman's "bagging" [7] and random selection of features [8], that operates by constructing some decision trees during the training process. The final prediction is an aggregation of the decisions made by trees in the forest. Majority vote is employed in the process of aggregation. The training process of random forest is divided into two parts: sampling randomly and splitting completely.

- Sampling randomly on records and features

The randomization of an RF is induced by bootstrap and random feature subspace. Bootstrap is used to select records randomly from the original datasets. A bootstrap sample set of size N is randomly drawn with replacement from an original dataset, which is also of size N [9]. So some samples may appear more than once after bootstrapping. Random feature subspace is used to randomly select features from the original feature space without replacement. The number of features selected is no more than the size of the original features from datasets. Sampling randomly on records and features makes it possible to train each tree in the forest with totally different samples, so that random forest could avoid overfitting well.

- Splitting Completely:

Each tree in forest is built without pruning [6]. In this way, a tree tends to be more different from the rest. The diversity of trees is another important reason to avoid overfitting well.

B. Random Forest on Imbalance Datasets

Imbalance degree of datasets probably affects the performance of random forest classifiers [10]. We assume that the negative effect may be caused by resampling with bootstrap in the random forest algorithm. Bootstrap means sampling with replacement, so for each sample in datasets, the probability of being selected is the same [11]. When it comes to imbalance datasets, bootstrap probably generates three kinds of sample sets: datasets with no minority samples, datasets with several minority samples but also larger imbalance coefficient than original datasets, datasets with some minority samples and the same or smaller imbalance coefficient compared with original datasets. However, the first and the second kind of sample sets are helpless. Because the decision trees trained based on these two kinds of sample sets could interfere with the final voting, which eventually leads to the degradation of the performance for random forest classifiers.

TABLE I. THE RELATIONSHIP BETWEEN AIR GRADES AND AIR QUALITY

Grades (Labels)	Air Quality Descriptions	Who needs to be concerned
1	Good	No need for concern
2	Moderate	Some people who may be unusually sensitive to ozone.
3	Unhealth for Sensitive Groups	Sensitive groups include: people with lung disease such as asthma, older adults, children and teenagers, and people who are active outdoors.
4	Unhealth	Everyone
5	Very Unhealth	Everyone
6	Hazardous	Everyone

TABLE II. THE IMBALANCE DEGREE OF AIR DATASETS

Dataset	Categories statistics		Imbalance coefficient
Dataset1	*1*	334	10.77
	2	733	
	3	412	
	4	251	
	5	133	
	6	46	
Dataset2	*1*	21443	1168.1
	2	12080	
	3	1266	
	4	135	
	5	39	
	6	6	
Dataset3	*1*	370	40.06
	2	204	
	3	30	
	4	10	
	5	5	
	6	2	
Dataset4	*1*	334	2.67
	2	334	
	3	320	
	4	318	
	5	318	
	6	285	

TABLE III. THE CONFIGURATIONS OF HYPER-PARAMETERS

Hyper parameter	Values	Description
criterion	"gini"	Gini impurity for datasets
n_estimators	178	The number of trees in the forest.
max_features	None	The number of features to consider when looking for the best split.
max_depth	None	The maximum depth of trees. None means nodes are expanded until all until all leaves contain less than *min_samples_split* samples
bootstrap	True (RF)	Whether bootstrap datasets are used when building trees.
sbootstrap	True (SGB-RF)	Whether samples grouped bootstrap datasets are used when building trees.

m*in_samples_split*	5	The minimum number of samples required to split an internal node.
min_samples_leaf	2	The minimum number of samples required to be at a leaf node

IV. THE PROPOSED ALGORITHM SGB-RF

In this section, we will introduce our modifications of the original RF algorithm, which adapts well to imbalance datasets. The improved RF algorithm is named Samples Grouped Bootstrap based Random Forest, SGB-RF in short.

The SGB-RF algorithm uses samples-grouped bootstrap method instead of bootstrap method in the random sampling phase. The samples-grouped bootstrap method is defined as follows:

- (Step1) Let C to be the number of categories in the origin sample set S. Group the original dataset S by class labels, samples with the same class label is in the same group. There are C new sample sets generated, which are different from each other with different class labels.
- (Step2)Use bootstrap method to resample in every sample set produced in step1 to generate another C bootstrap sample sets.
- (Step3)Merge C sample sets generated in step2 to a new sample set denoted as S_g, which size is the same with S.

Samples-grouped bootstrap method guarantees the randomness while maintaining the imbalance degree of new sample sets S_g. On the one hand, samples-grouped bootstrap inherits the schema of bootstrap while sampling in subsamples with the same label. On the other hand, the imbalance coefficient of new datasets S_g is equals to the imbalance coefficient of the original dataset S.

V. EXPERIMENTS AND ANALYSIS

In order to assess the performance of the proposed method, we design three sets of experiments based on datasets in TABLE II. In the experiments, the decision trees are constructed by using CART algorithm which uses the Gini index to measure purity of nodes and uses the minimum distance based Gini index to select the splitting attribute [12]. In the following experiments, some configurable parameters are configured as TABLE III, which we chose the best by experimenting with random search strategy.

A. Experiment (1)

We train a set of RF classifiers based on *dataset1* and train another set of RF classifiers based on *dataset4*. Both of them make predictions on *dataset3*. Prediction classification reports of the two sets of RF classifiers are presented in TABLE IV and TABLE V. According to values of precision, recall, and F1 from classification reports, RF classifiers based on *dataset4* perform better than RF classifiers based on *dataset1* of 2 percentage points. From TABLE II, the imbalance coefficient of *dataset1* is 10.77, while the imbalance coefficient of *dataset4* is 2.67. The results of this experiment prove that the performance of RF classifiers is related to the imbalance degree of datasets. RF classifiers, trained on imbalance datasets, usually perform unsatisfactorily.

TABLE IV. THE CLASSIFICATION REPORT OF RF CLASSIFIERS BASED ON DATASET1

Labels	Classification Report			
	Precision	***Recall***	***F1 - score***	***Support***
1	1.00	0.89	0.94	370
2	0.83	0.92	0.87	204
3	0.64	0.93	0.76	30
4	0.82	1.00	0.90	9
5	1.00	1.00	1.00	5
6	0.00	0.00	0.00	2
avg / total	0.90	0.89	0.89	620

TABLE V. THE CLASSIFICATION REPORT OF RF CLASSIFIERS BASED ON DATASET4

Labels	Classification Report			
	Precision	***Recall***	***F1 - score***	***Support***
1	0.99	0.91	0.95	370
2	0.84	0.90	0.87	204
3	0.60	1.00	0.75	30
4	1.00	0.89	0.94	9
5	1.00	1.00	1.00	5
6	1.00	1.00	1.00	2
avg / total	0.92	0.91	0.91	620

B. Experiment (2)

This experiment is designed to verify that the SGB-RF algorithm could improve the performance of classification on imbalance datasets compared with RF. Training samples come from *dataset2*, while the testing samples come from *dataset1*. RF classifiers and SGB-RF classifiers are trained with the configuration of hyper-parameters showed in TABLE III.

We use confusion matrix to evaluate the classification performance. Confusion matrix is a very effective way to assess the accuracy of imbalance datasets, in that the accuracy of each category are plainly described along with both the errors of inclusion and errors of exclusion present in the classification [13]. Fig. 1, Fig. 2 are the confusion matrix of RF classifiers and SGB-RF classifiers tested on *dataset1*. For the minority sample set, which size is 46 and class label is *"6"*, all samples are misclassified in RF classifiers while 16 of 46 samples are classified correctly by SGB-RF classifiers. For samples whose label is *"3"* and *"5"*, the accuracy of SGB-RF classifiers is higher than the accuracy of RF classifiers. For the rest samples labeled as *"1", "2"* and *"4"*, RF classifiers and SGB-RF classifiers get the same performance. The results of this experiment prove that SGB-RF performs better than RF, especially in the classification on minority samples.

C. Experiment (3)

For further verify the effectiveness of SGB-RF, we also evaluate the proposed method on 4 datasets from UCI repository [14]. TABLE VI shows the description of datasets used in this experiment, including imbalance coefficient to measure the imbalance degree of datasets. According to the imbalance coefficient, *Glass* is the most unbalanced, followed by *Wine>Breast>Iris*.

Breiman has given empirical evidence in [15] to show that the out-of-bag estimate is as accurate as using a test set of the same size as the training set. So in this experiment we evaluate the performance of SGB-RF classifiers and RF classifiers with out-of-bag scores. The higher the score is, the better a model is. Fig.3 shows the results obtained by this experiment. It is shown that SGB-RF presents an improvement(about 1 percentage point)

of RF when both apply on balance dataset *Iris*. The improvement is significant when they apply on imbalance datasets(up to 6 percentage points on *Glass*).

VI. Conclusion

Air pollution is a serious problem for the health of residents, so it's important to make an accurate evaluation on the grades of air quality as soon as possible. Grades of air quality help residents take measures in advance to reduce the effects of air pollution on health. In this paper, we find that the distribution of air data is imbalance, which leads to the decrease of performance from random forest classifiers. In order to adapt to the imbalance datasets, we propose a samples grouped bootstrap based random forest algorithm(SGB-RF). Through three sets of experiments, we make the following observation: the performance of random forest classifiers perform unsatisfactorily on imbalance datasets. SGB-RF presents an improvement of RF when both apply on balance datasets. The improvement is significant when they apply on imbalance datasets, where SGB-RF is much better than RF for the strong ability to classify minority samples correctly.

For future research, we will compare the proposed algorithm with ensemble methods based on boosting, and the proposed algorithm will be further explored for more extensive applications on other datasets.

Acknowledgment

This work is supported by The Major Science & Technology Program of Guangxi (Grant No.GKAA17129002) and The Key Research Program of Chongqing Science & Technology Commission (Grant No.CSTC2017jcyjBX0025).

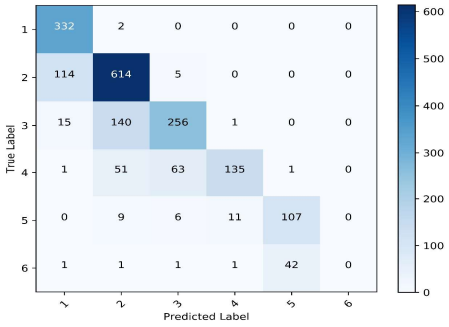

Fig. 1. The confusion matrix of RF classifiers

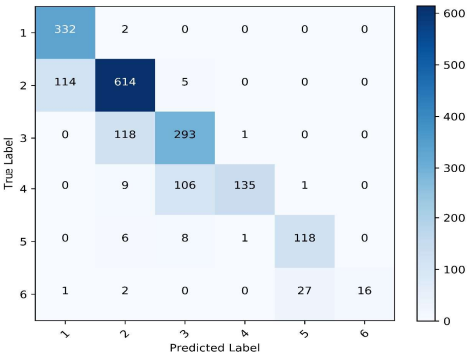

Fig. 2. The confusion matrix of SGB-RF classifiers

TABLE VI. The Description of UCI Datasets Used in Experiment

Dataset	Number of records	Number of features	Number of categories	Imbalance coefficient
Breast	116	9	2	1.39
Glass	214	9	6	8.99
Iris	150	4	3	1
Wine	178	13	3	3.91

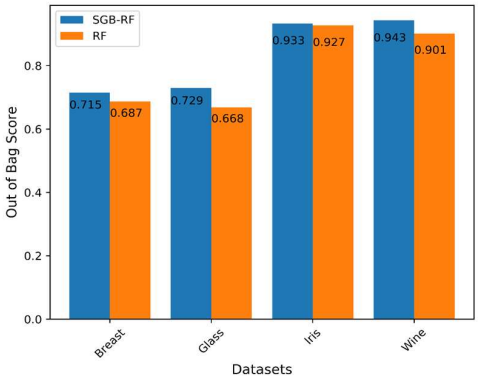

Fig. 3. The out-of-bag scores of RF and SGB-RF on UCI datasets

References

[1] Sarkheil, H. & Rahbari, S. Environ Earth Sci (2016) 75: 1319. https://doi.org/10.1007/s12665-016-6131-2

[2] Yun Bai, Yong Li, Xiaoxue Wang, Jingjing Xie, Chuan Li,Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions,Atmospheric Pollution Research,Volume 7, Issue 3,2016,Pages 557-566

[3] A. Shawabkeh, F. Al-Beqain, A. Redan and M. Salem, "Benzene Air Pollution Monitoring Model using ANN and SVM," 2018 Fifth HCT Information Technology Trends (ITT), Dubai, United Arab Emirates, 2018, pp. 197-204

[4] Sagi, O, Rokach, L. Ensemble learning: A survey. WIREs Data Mining Knowl Discov. 2018; 8:e1249. https://doi.org/10.1002/widm.1249

[5] [Online] Avaliable: https://www.cnemc.cn/

[6] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. doi:10.1023/A:1018054314350

[7] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[8] Carmen Lai, Marcel J.T. Reinders, Lodewyk Wessels,Random subspace method for multivariate feature selection,Pattern Recognition Letters,Volume 27, Issue 10,2006,Pages 1067-1076

[9] Echeverri, A.C., von Harling, B. & Serone, M. J. High Energ. Phys. (2016) 2016: 97. https://doi.org/10.1007/JHEP09(2016)097

[10] Salvador García, Zhong-Liang Zhang, Abdulrahman Altalhi, Saleh Alshomrani, Francisco Herrera,Dynamic ensemble selection for multi-class imbalanced datasets,Information Sciences,Volumes 445–446,2018, Pages 22-37

[11] José A. Sáez, Bartosz Krawczyk, Michał Woźniak,Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets,Pattern Recognition,Volume 57,2016,Pages 164-178. https://doi.org/10.1016/j.patcog.2016.03.012

[12] Breiman L, Friedman J, Stone C. Classification and Regression Trees. Wasworth, 1984

[13] Caelen, O. Ann Math Artif Intell (2017) 81: 429. https://doi.org/10.1007/s10472-017-9564-8

[14] [Online] Avaliable: http://archive.ics.uci.edu/ml/

[15] Breiman, L. (1996) Out-of-Bag Estimation. https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf