IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 2, 2019, accepted November 26, 2019, date of publication November 28, 2019,
date of current version December 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2956599

Air Quality Predictive Modeling Based on an Improved Decision Tree in a Weather-Smart Grid

YUANNI WANG[1,2] **AND TAO KONG**[1]

[1]School of Computer Science, China University of Geosciences, Wuhan 430078, China
[2]Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430078, China

Corresponding author: Yuanni Wang (ynwang2005@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 41773063, and in part by the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing under Grant KLIGIP-2018B04.

ABSTRACT With various developments, the concept of the smart city has attracted great attention all over the world. To many, it is a good intelligent response to the needs of people's livelihoods, environmental protection, public safety, etc. A weather-smart grid is an important component of the smart city, and the health of the weather-smart grid will directly affect the health of the smart city. Efficient and accurate predictions about air quality levels can provide a reliable basis for societal decisions, safety for smart transportation, and weather-related disaster prevention and preparation. To improve the time performance and accuracy of prediction with a large amount of data, this paper proposes an improved decision tree method. Based on an existing method, the model is improved in two aspects: the feature attribute value and the weighting of the information gain. Both accuracy and computational complexity are improved. The experimental results demonstrate that the improved model has great advantages in terms of the accuracy and computational complexity compared with the traditional methods. Additionally, it is more efficient in addressing classification and prediction with a large amount of air quality data. Moreover, it has good prediction ability for future data.

INDEX TERMS Smart city, weather-smart grid, predictive modelling, air quality, decision tree, discretization, weighting.

I. INTRODUCTION

Smart city design involves many aspects of the urban ecological environment, including a weather-smart grid, transportation, medical treatment, intelligent buildings, etc. [1]. A weather-smart grid is essential for a smart city. A weather-smart grid applies big data, machine learning and artificial intelligence technology to make more accurate weather forecasts. It improves the overall understanding of the environment, the ability to prepare for extreme weather, climate and water events and provides decision support for disaster reduction and prevention. A weather-smart grid has a large impact on a smart city [2]. The use of cloud computing, the internet of things (IoT), mobile interconnections, big data and intelligent control technologies to keep weather forecasts accurate and intelligent needs to be focused.

Urban air pollution is a common problem. It seriously affects human health, and is accompanied by the emergence of various diseases, such as lung cancer [3], [4]. Some other serious environmental problems may also be caused by air pollution, such as acid rain and greenhouse gas effects. Thus, air pollution is currently one of the most alarming concerns.

Currently, high-precision air sensors are mostly used to monitor air quality. This allows for the collection of more accurate data. From these data, air prediction and analysis can be generated in advance for scientific decision support as well as clustering and classification [5]–[7].

The problem of air pollution can be fundamentally addressed only by taking effective measures to prevent and control pollution before it occurs. Therefore, it is very important to establish an effective model for evaluating and predicting air quality in a timely manner.

According to the air quality index (AQI) used in China, ambient air pollutants are concentrations of particulate matter (PM10 and PM2.5), sulphur dioxide (SO_2), nitric oxide (NO), nitrogen dioxide (NO_2) and other nitrogen oxides (NOx). Previously, the traditionally derived AQI was used to predict the degree of air pollution. According to the proportion of

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Luo.

various components in the air, the monitored air concentration was simplified to a single conceptual index. It classified the degree of air pollution and air quality status and was suitable for expressing the short-term air quality status and the changing trends in a city [8], [9]. With advancements in technology and research, many alternative methods have been proposed that use big-data and machine-learning approaches [10]. However, these air quality models are limited by the computational costs. Traditional machine-learning methods focus on the accuracy of classification, which makes it difficult to reduce the cost of calculation. To improve the time performance, a novel algorithm is needed.

Among all techniques, machine-learning algorithms are the most widely used to for classification of air quality evaluation. Muhammed et al. noted that machine-learning algorithms were best suited for air quality prediction [11]. With the rise of artificial intelligence in recent years, traditional machine learning and deep learning have also been successfully applied in the field of air quality prediction and have achieved good results [10]–[14]. For example, the decision tree method has been used with common decision tree algorithms, such as ID3, C4.5 and CART [15], [16]. In addition, the k-nearest neighbour algorithm has also been used [17], [18]. Later, with the development of ensemble learning, random forests were also applied in this field [19], [20]. Random forests followed a technique as per [21], where several decision trees were built based on subsets of data, and an aggregation of the predictions was used as the final prediction. In recent years, big data, neural networks and in-depth learning have been gradually applied to this field. For example, the earliest artificial neural network (ANN) was used to process time-series data, and then, a recursive neural network (RNN) was used to predict future air quality changes by using concentration changes in NO_2, CO_2, SO_2, PM2.5 and other air pollutants from the previous period [22], [23]. Later, a long-term and short-term memory model (LSTM) was developed, which could retain longer time intervals and eliminated the problem of partial gradient disappearance. This model had good application prospects for air quality prediction [24]. Almost all the related studies focused on achieving better accuracy, but little attention was paid to the complexity of the algorithms.

In this paper, the computational cost and execution efficiency of the algorithm are considered in the case of a large amount of data. This paper tries to reduce the computational cost and execution efficiency of the algorithm as much as possible while guaranteeing a certain accuracy. Hence, a novel algorithm of improved decision tree C4.5 is proposed for air quality prediction. The model is improved in two aspects: the feature attribute value and the weighting of the information gain. Both accuracy and computational complexity are improved. It is based on the attribute partition principle of the information gain rate of C4.5. To improve the efficiency of the algorithm execution as much as possible, we discretize the data before computing the information entropy. The information gain rate of N attribute values at the boundary points is calculated, and the attribute value with the greatest information gain is selected as the optimal segmentation threshold to divide it, replacing the information gain rate obtained by traversing all the attribute values in the traditional C4.5 algorithm. In addition, a weighted coefficient w is introduced to calculate the increment rate of each attribute to consider the extent of the impact of various pollutants on the air and improve the accuracy of the new algorithm. The indicators used for evaluating the classification prediction models are the receiver operating characteristic (ROC) curve, precision-recall (PR) curve, confusion matrix, etc. The improved algorithm has great advantages in accuracy and computational complexity. It is more efficient in dealing with classification and prediction given a large amount of air quality data.

The paper is organized as follows. In Section II the related work is discussed including the issue of using other algorithms. Then, the new improved algorithm is introduced in detail in Section III. The experimental settings are described in Section IV. Section V provides the analysis and experimental results. We conclude with the contribution of our work in Section VI.

II. RELATED WORK

Air quality evaluation is an important way to monitor and control air pollution. Faced with a deteriorating atmospheric environment, there are many research papers that focus on classification in air quality evaluation. Many various methods are constantly derived, and many achievements have been made in the efficiency and accuracy of the related algorithms.

In the field of air quality evaluation, there are many prediction methods. The most direct and efficient method is to calculate the air pollution index (API) according to the concentration value of designated air pollutants. Air quality is directly obtained through the analysis of the pollution index. This method is suitable for the analysis of air quality status and changing trends in the short term [25].

Air quality evaluation has been conducted using conventional approaches for many years. The traditional approaches for air quality prediction use mathematical and statistical techniques [26]. In these techniques, a physical model is initially designed, and data is coded with mathematical equations. However, these are complex mathematical calculations. In addition, these methods provide limited accuracy. More recently, alternatives to traditional methods have been proposed; these alternatives use big-data and machine-learning approaches [10]. Many researchers have developed or used big-data-analytical models and machine-learning-based models to conduct air quality evaluation to achieve better accuracy in evaluation and prediction. There are several common algorithms used for air quality evaluation, such as the air quality index method, clustering analysis, the artificial neural network model, the decision tree model, the random forests model, the least squares support vector machine model and the deep belief network.

The clustering algorithm is simple and not difficult to implement, but its stability is poor, and it is very sensitive

to the selection of initial clustering centres. The clustering process often has difficulty in converging due to the improper selection of clustering centres and can fall into a situation of a dead cycle or a local optimum, which was unacceptable to us. When the sample data are large and the distribution of attribute values is complex, the clustering centres need to be updated frequently; thus, the time efficiency of the algorithm is low. Based on these shortcomings, clustering analysis may not be a good choice for air quality evaluation.

With the rise of artificial intelligence in recent years, traditional machine learning and deep learning have also been successfully applied in the field of air quality evaluation and have achieved good results [9], [10]. For example, the air quality data are extracted and the information gain or gain rate is calculated. Then, the optimal attribute partition point is selected to establish the decision tree for classification. Common decision tree algorithms, such as ID3, C4.5 and CART, are used. Decision trees are commonly used in air quality evaluation. Currently, many researchers propose a variety of decision tree algorithms and have extended their application in a variety of fields. In air quality prediction, the research results regarding the application of decision trees have been impressive [27]–[30]. With the increase in the acquisition of air quality data, the value of attributes is also increasing. Some problems will arise when using decision trees to evaluate air quality. As the number of nodes increases, the number of continuous attributes increases, and the value of any attribute in the continuous attributes increases; these increases will have a negative impact as frequent logarithmic operations in the process of calculating information gain seriously affect the performance of the algorithm, which will greatly affect the efficiency of the decision tree generation.

With the development of ensemble learning, the random forest has also been applied in this field. N decision tree classification models of the same or different types are integrated together to form a random forest. Each sub-tree of the forest corresponds to a decision tree classifier [31].

With the development of big data, the amount of data is also increasing. Neural networks and deep learning have also been gradually applied to this field. For example, artificial neural networks were introduced for examining the concentration of various air pollutants. Through the calculation of multiple perceptron and excitation functions, the air quality grade was finally classified [32]–[34]. Then, there was the emergence of the recursive neural network (RNN), which connects and trains multiple layers of neurons by the establishment of weights on the connections. This network can be used to process time-series data and predict future air quality changes by using the changes in air pollutant concentrations (such as NO_2, CO_2, SO_2 and PM2.5) in the previous period [22], [23]. When there are more layers in the RNN, the gradient will disappear. Therefore, researchers put forward the long short-term memory (LSTM) model, which can remember longer time intervals and eliminates the problem of partial gradient disappearance. This model has good application prospects for air quality prediction [35]. These methods have achieved good results in accuracy, but they do not provide adequate consideration of the time performance of the algorithm.

III. MATHEMATICAL MODEL

A. DECISION TREE C4.5

The ID3 algorithm is the most well-known decision tree algorithm. It is based on information theory. The core idea is to use the information gain as a measure of attribute selection. The C4.5 algorithm improves it by using the information gain rate to select node attributes, which mainly overcomes the shortcomings of the ID3 algorithm in choosing attributes with more values.

The main ideas of the C4.5 algorithm are as follows:

Assuming that S is a training set, the target attribute C of S has m possible class label values, $C = \{C_1, C_2, \ldots C_m\}$. In training set S, the frequency of C_i in all samples is $p_i (I = 1, 2, 3, \ldots, m)$, and then the information entropy contained in the training set S is defined as:

$$Entropy(S) = Entropy(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{1}$$

Assuming attribute A is used to partition S, the information entropy of the partitioned sample subset is as follows:

$$Entropy_A(S) = \sum_{i=1}^{k} \frac{|s_i|}{|s|} Entropy(S_i) \tag{2}$$

Assume that there are v possible values $\{a^1, a^2, \ldots, a^v\}$ of continuous attribute a. If a is used to partition the sample set S, v branch nodes will be generated. The vth branch node contains a sample whose value is a^v on all attributes a in S, which is denoted as S^v. Then, the information entropy of S^v, $Entropy(S^v)$, can be obtained. Considering that the number of samples contained in different branch nodes is different, the information gain obtained by partitioning the sample set S with attribute a can be calculated.

$$Gain(S, a) = Entropy(S) - \sum_{v=1}^{v} \frac{|S^v|}{|S|} Entropy(S^v) \tag{3}$$

Because the information gain criterion has a preference for attributes with more desirable values, C4.5 does not use the information gain directly but uses the information gain ratio to select the best partitioned attribute value to reduce the adverse effects of the preference.

The information gain ratio normalizes the information gain by using the split information value. The definitions are as follows:

$$SplitInfo_A(S, a) = -\sum_{j=1}^{v} \frac{|s_j|}{|s|} \times \log_2 (\frac{|s_j|}{|s|}) \tag{4}$$

This value represents the information generated by dividing the training data set S into v partitions corresponding

to the v outputs of the attribute a test. The definition of the information gain ratio is:

$$GainRatio(S, a) = \frac{Gain(S, a)}{SplitInfo(S, a)} \tag{5}$$

The attributes with a maximum gain rate are selected as the splitting attributes.

B. NEW ALGORITHMIC THOUGHT

C4.5 enhances the function of the ID3 algorithm, but there are still many shortcomings in the detection of air quality using the C4.5 algorithm. As the amount of air quality data increases, the value of attributes also increases. According to the Taylor expansion theorem, we know that logarithmic operations are more complex. Therefore, in the process of calculating the information gain, frequent logarithmic operations will seriously affect the performance of the algorithm.

When discretizing continuous attributes, the C4.5 algorithm inserts several segmentation points into different values of any attribute, calculates the information gain rate of all segmentation points and chooses the segmentation threshold with the largest information gain rate as the best segmentation threshold of the continuous attributes. When the number of nodes in the decision tree is large, the number of continuous attributes is large and the value of any attribute in the continuous attributes is large, the computational complexity of the algorithm is considerable, which will greatly affect the efficiency of the decision tree generation.

Using the traditional C4.5 algorithm to predict air quality data will neglect some influence of the data itself on the results; it will simplify the model in calculating the information gain rate, but without consideration of the extent of the impact of various pollutants on the air. Therefore, we need to weigh the gain rate of each attribute, reduce the information entropy of some attributes and provide the information entropy of other attributes accordingly.

C. IMPROVED C4.5 ALGORITHM

The procedure for the improved C4.5 algorithm is shown in the following figure.

After data pretreatment and discretization, a few attribute values, which are candidate partition points, are obtained and then classified, tested and output by the classification model.

The improved decision tree uses the attribute value with the largest gain rate of weighted information as the partition node, generates new branches and then establishes the branches of the decision tree nodes by recursive invocation of this method.

The algorithm in this paper can be summarized in the following seven steps:

(1) *Read the text data and save the attributes of the air data and the corresponding pollution level.*

(2) *Clean the data. Handle outliers and defaults.*

In some cases, the data provided for use may lack the values of certain attributes. In this case, it is often necessary to estimate the missing attribute values based on other instances

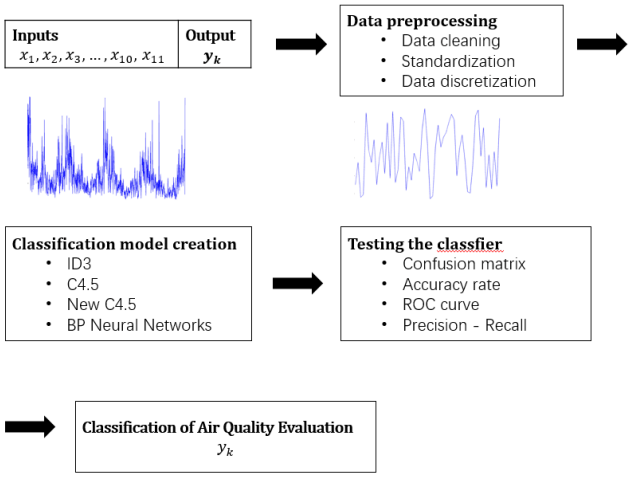

FIGURE 1. The procedure flow chart of the new C4.5 algorithm.

where the attribute value is known. In this paper, the method of average estimation is used to address the default values. If a certain attribute value of the sample data is missing, the average of the attribute values of all the data of the corresponding category is calculated, and the average value is used to estimate the default value. The specific calculation methods are as follows:

Assume the sample set is S, and $S_{i,j}$ is the value of attribute j of the ith sample. If this value is missing, the estimated value $E_{i,j}$ is calculated as follows:

$$E_{i,j} = \frac{\sum_{k=1}^{m} S_{k,j}}{m}, i = 1, 2, \ldots n \tag{6}$$

$S_{k,j}$ represents the same corresponding attribute values as those of the ith sample label. M represents the amount of the same class as that of sample I.

(3) *Discretize the data values under each attribute.*

In the process of discretization of continuous attributes, the C4.5 algorithm needs to predict all the partitions, which requires much time. How to select an optimal partition threshold quickly has become an urgent problem to be solved. In this paper, we use the following methods to discretize.

(i) Finding the boundary point value

Fayyad et al. proved that no matter how many classes are used in the data set for learning, and no matter how the classes are distributed, the best segmentation points of continuous attributes always lay at the boundary points. According to the Fayyad boundary point principle, continuous descriptive attributes are arranged in ascending order. Assuming the number of data attributes is n, the values of the n attributes at adjacent class boundary points of a continuous attribute point are $a_1, a_2, \ldots a_n$, in which the test attribute value a_i is the maximum value in class i. The corresponding information gain is then calculated, and the attribute value with the largest information gain is selected as the optimal segmentation threshold.

(ii) Interpolation

Given sample set S and continuous attribute a, assuming that a has n different values on S, these values are ranked from smallest to largest and marked as $\{a_1, a_2, \ldots, a_n\}$. Based on the partition point t, S can be divided into S_t^+. Among them, S_t^- contains samples whose S_t^- values are not greater than t on attribute a, while S_t^+ contains samples whose values are greater than t on attribute a.

Because each set of boundary points is determined by the number of classes, when there are few classes, the number of boundary points is very small, which will lead to low accuracy of the model and is not conducive to the generalizability of the algorithm. Therefore, we need to interpolate the boundary points appropriately to improve the accuracy of the model.

The improved algorithm can obtain a new set of candidate partition points labelled T_a, where $\{a_1, a_2, \ldots a_m\}$ $(m > n)$, by searching for boundary points and interpolation operations. It needs to calculate only the information gain rate of N attribute values at the boundary points, in contrast to the traditional C4.5 algorithm, which traverses all the attribute values. The computational complexity is greatly reduced, and it will not increase much when the amount of data increases. When there is a special case in which each attribute value represents only one category, the computational complexity of the improved algorithm is comparable to that of the C4.5 algorithm.

(4) *Calculate the weighted information gain rate of each attribute value after discretization.*

First, we need to calculate the information entropy. Here, we replace the log operation with a Taylor expansion to reduce the amount of calculation.

Taylor's mean value theorem: If there are $n+1$ derivatives of $f(x)$ in an open interval (a, b) including x_0, then when $x \in (a, b)$, there is:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(\zeta)}{2!}(x - x_0)^2 (x \le \zeta \le x) \tag{7}$$

The operation of $\log_2(x)$ is expanded as above, with $f(x) = \log_2(x)$.

$$\log_2(x) = \frac{In(x)}{In2} \tag{8}$$

$$In(x) \approx In(x_0) + \frac{1}{x_0}(x - x_0) - \frac{1}{2! \times x_0^2}(x - x_0)^2 \tag{9}$$

Setting $x_0 = 1$, we obtain the following:

$$In(x) \approx \frac{2(x-1) - (x-1)^2}{2} \tag{10}$$

$$log_2(x) \approx \frac{2(x-1) - (x-1)^2}{2In(2)} \tag{11}$$

Therefore, the calculation of the information entropy is as follows:

$$Entropy(S) \approx -\sum_{i=1}^{m} p_i * \frac{2(p_i - 1) - (p_i - 1)^2}{2In(2)} \tag{12}$$

A weighting coefficient w $(0 < w < 1)$ is introduced, which is determined by the decision maker according to prior knowledge. By weighting, we can reduce the information entropy of some attributes, and correspondingly, improve the information entropy of other attributes.

If the weighting coefficient of an attribute is w, the information gain ratio is calculated as:

$$Entropy(S_t)' = mean(Entropy(S_t^{\lambda}))\lambda \in \{-, +\}, \quad t \in T_a \tag{13}$$

Here, we refer to the information entropy. S is a given sample set. a is a continuous property. T_a is the set of attribute partition points after the discretization operation. We classify each partition point t into two categories: positive and negative. λ represents the classification.

$$w_{S_t} = \frac{1 - Entropy(S_t)'}{m - \sum_{t=1}^{m} Entropy(S_t)'} \tag{14}$$

The weights are calculated relative to each partition point according to the above formula. m represents the number of partition points. The bigger the information entropy is, the smaller the weight is.

$$Gain(S, a, t) = w_{S_t} \times (Entropy(S) - \sum\nolimits_{\lambda \in \{-,+\}} \frac{|S_t^{\lambda}|}{|S|} Entropy(S_t^{\lambda})) \tag{15}$$

Gain is the weighted information gain. We adjust the value of the information gain appropriately by a weighting operation to achieve a better effect of the model.

The information gain ratio is normalized by using the split information value, which is defined as follows:

$$SplitInfo_A(S, a) = -\sum_{j=1}^{v} \frac{|s_j|}{|s|} \times \log_2(\frac{|s_j|}{|s|}) \tag{16}$$

This value represents the information generated by dividing the training data set S into v partitions corresponding to the V outputs of the attribute a test. V corresponds to the number of partition points corresponding to the attribute. We divide each partition point into two categories. V partition points yield V partition results; subsequently, V information entropy is obtained and then summed.

$$GainRatio(S, a, t) = \frac{Gain(S, a, t)}{SplitInfo(S, a)} \tag{17}$$

$$GainRatio(S, a) = max_{t \in T_a} GainRatio(S, a, t) \tag{18}$$

Equation 18 is the information gain rate obtained.

(5) *The attributes with the maximum weighted information gain rate are selected as the root nodes of the subtree.*

(6) *If the candidate attribute value is not empty, the root node generates a new branch node.*

(7) *The branch process is carried out recursively until the candidate node is empty.*

IV. EXPERIMENTAL SETTINGS

A. DATASET GENERATION

For the same method, to test the adaptability of the algorithm, three experiments were designed. It was from two aspects: the change of data quantity and the prediction ability of future data. The the data variation mainly included general data, with missing value and a large amount of data. Experiment 1 had a general amount of data and contained missing values. It was of universal significance to verify the effectiveness of the algorithm. Experiment 2 had a large amount of data and no missing values. It was used to verify the time efficiency and performance of the algorithm in the case of a large amount of data. Experiment 2 illustrated the advantages of the algorithm. The training data and prediction data of Experiment 1 and Experiment 2 were randomly selected and allocated from the historical data, only the unknown samples were predicted from the known samples, and the classification by time was not considered. To test the new algorithm for the prediction of future data, Experiment 3 was designed. Compared with experiment 2, the training data and test data were classified according to the time sequence, which was more practical for the prediction of future data.

(i) Data for Experiment 1

The air quality data of a city were selected [36]. The data consisted of 320 records. There were default values in some attributes of the data.

There were a few missing values in the collected data, so before data mining, the default values were processed according to the method of average estimation to fill those missing values.

The data was divided into training and test sets with a 7:3 ratio.

(ii) Data for Experiment 2

The experimental data were obtained from a query of the Chinese air quality historical data website [37]. For the experiment, we selected nearly 1970 data records, in which the training and test sets were generated with a distribution ratio of 7:3 (1372 in the training set, 588 in the test set).

(iii) Data for Experiment 3

Data from 2014-2017 for the city of Wuhan were selected as the training set and data from 2018 were used as the testing set.

B. SETTING OF EXPERIMENTAL PARAMETERS

Weight calculation modification was involved in the experimental process. In constructing the decision tree, the split feature points were selected by the weighted information gain ratio, and the weights were expressed by w. The weights of w were calculated according to Step 4 in Section III-C.

C. RESULT EVALUATION METRICS

Computing a meaningful estimate of generalizability is a key requirement for evaluating the performance of a classification algorithm [38].

Three metrics were introduced to evaluate the performance of the classification results in this paper: the ROC curve [39], [40], the PR curve, and the confusion matrix. These metrics are described in the following.

The first performance metric that was chosen is the ROC curve [39]–[41].

The ROC curve relates the true positive rate (TPR) to the false positive rate (FPR) obtained at every possible threshold. Thus, it provides insight into the performance of a classifier that is independent of its detection threshold.

True positives (TPs) correspond to the number of correct matches found, while false negatives (FNs) represent the number of correct matches not found. False positives (FPs) denote the number of non-matches incorrectly identified as matches, i.e., misinformation. True negatives (TNs) indicate the number of mismatches rejected correctly. From these, one can calculate the true positive rate (TPR), the false positive rate (FPR) and the true negative rate (TNR).

$$TPR = \frac{TP}{TP + FN} \tag{19}$$

$$FPR = \frac{FP}{FP + TN} \tag{20}$$

$$TNR = \frac{TN}{FP + TN} \tag{21}$$

In the ROC curve, the transverse axis is the FPR, and the larger the FPR, the more actual negative classes are predicted as part of the positive classes. The longitudinal axis is the TPR. The larger the TPR, the more actual positive classes are predicted. The closer the ROC curve is to (0, 1), the better it is to deviate from the 45-degree diagonal line. The greater the TPR and TNR values are, the better the effect is.

The second metric is the PR curve. Assume that P is accuracy and R is the recall rate. TP, FP, FN, and TN are as described above for the ROC curve.

$$P = \frac{TP}{TP + FP} \tag{22}$$

$$R = \frac{TP}{TP + FN} \tag{23}$$

The third performance metric we have chosen is the confusion matrix [42].

Assume that for the classification task of N-type patterns, the recognition data set D includes T_0 samples, and each type of pattern contains T_i data. Classifier C is constructed by some recognition algorithm. cm_{ij} represents the percentage of the data classified by classifier C into class j patterns in the total number of samples of class i patterns. The following N^*N-dimensional confusion matrix CM (C, D) can be obtained.

$$CM(C, D) = \begin{pmatrix} cm_{11} \ldots\ldots cm_{1N} \\ cm_{21} \ldots\ldots cm_{2N} \\ \ldots\ldots\ldots\ldots\ldots \\ cm_{N1} \ldots\ldots cm_{NN} \end{pmatrix}$$

The correct recognition rate of each pattern is described as follows:

$$R_i = cm_{ii}, \quad i = 1, \ldots, N \tag{24}$$

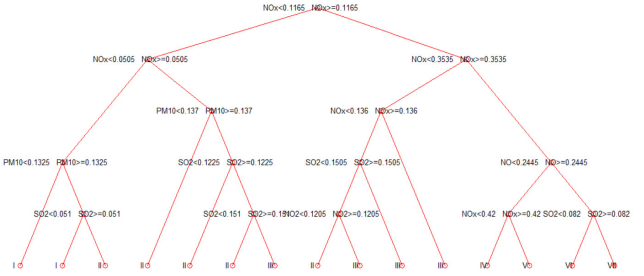

FIGURE 2. Decision tree generation of experiment 1.

The average correct recognition rate is described as follows:

$$R_A = \sum_{i=1}^{N} (cm_{ii} * T_i) / T_0 \tag{25}$$

The error recognition rate of each pattern is described as follows:

$$W_i = \sum_{j=1, j\neq i}^{N} cm_{ij} = 1 - cm_{ii} = 1 - R_i \tag{26}$$

The average error recognition rate is described as follows:

$$W_A = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (cm_{ii} * T_i)/T_0 = 1 - R_A \tag{27}$$

V. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results are presented in this section. To assess the performance of the new improved C4.5, the ROC curve, PR curve, run-time and accuracy were applied. We compared the results from three aspects: (1) the performance of different classification methods; (2) the time and accuracy of different classification methods (the results are described in detail); and (3) the prediction of future air quality from historical data.

All experiments were carried out on a desktop computer with the following configuration: CPU (Intel-i5-4210H); RAM (8GB); operating system (Windows 10 Professional). The experiments were executed on MATLAB 2013a.

A. RESULTS OF THE EXPERIMENTS

In these experiments, the highest height of the decision tree was set to 6, the category was the air quality grade, and the number of categories was 6. The six grades were *excellent*, *good*, *mild*, *moderate*, *severe* and *severe in turn*, expressed as I, II, III, IV, V and VI, respectively.

Here, we give the result of the decision tree in experiment 1.

B. PERFORMANCE COMPARISON OF THE DIFFERENT ALGORITHMS

The data of Experiment 1 were general. There was a large amount of data in Experiment 2, which was representative of too much data processing. To evaluate the prediction model of this algorithm, the ROC curve and PR curve were selected

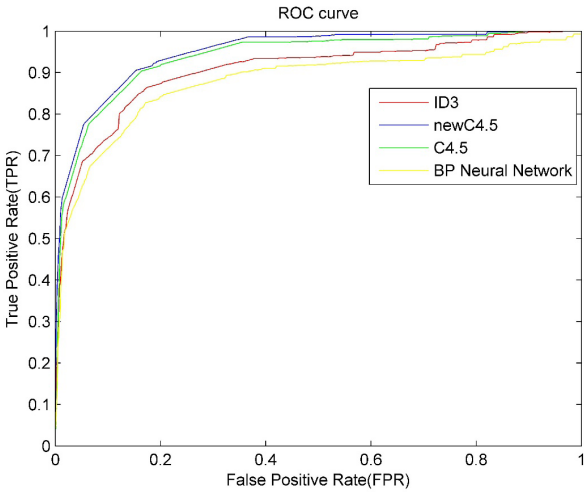

FIGURE 3. ROC curve of experiment 2.

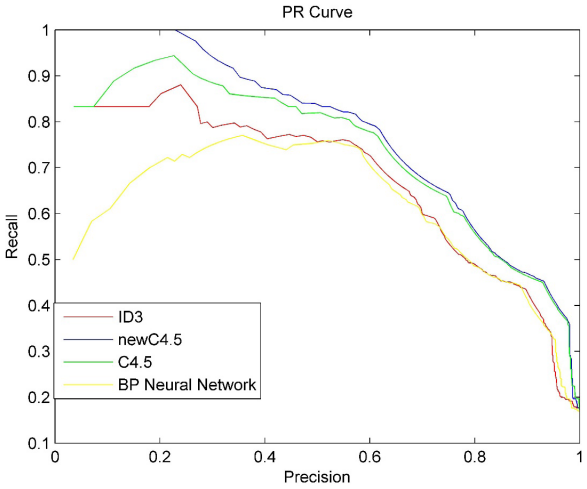

FIGURE 4. PR curve of experiment 2.

to analyse the performance of the model. The results of Experiment 2 are shown in Figures 3 and 4. The algorithm was compared with ID3, C4.5 and a backpropagation (BP) neural network.

From the ROC curve of Figure 3, we can see that the curve of the new C4.5 model was convex at the far left. The true positive parameter was higher, the false positive parameter was lower, and the AUC (area under the curve) was larger under the ROC curve. We can see that the AUC was the largest under the new C4.5 model curve. From the PR curve of Figure 4, we can see that the PR curve corresponding to the new C4.5 model was convex to the far right and closest to the coordinates (1, 1). The precision and recall were both high, indicating that the new C4.5 model was the best.

C. RUN-TIME AND ACCURACY COMPARISON

In addition, the time spent constructing the decision tree was greatly reduced while the accuracy of the algorithm was guaranteed. Tables 1 and 2 below show the accuracy and run-time of the algorithm and compare them with the performances of ID3, C4.5 and the BP neural network.

TABLE 1. Accuracy and run-time comparison on test sets of experiment 1.

Algorithm	Accuracy(%)	Run-time (s)
ID3	92%	0.437556
New C4.5	98%	0.061435
C4.5	95%	0.438262
BP Neural Network	91%	1.060178

TABLE 2. Accuracy and run-time comparison on test sets of experiment 2.

Algorithm	Accuracy(%)	Run-time (s)
ID3	91.6%	2.481025
New C4.5	96.6%	0.753725
C4.5	95.5%	1.24291
BP Neural Network	89.8%	4.588766

From Table 1, we can see that the run-time of the new C 4.5 method was approximately 10 times shorter than that of the other methods. From Table 2, we can also see that when the amount of data was large, the run-time of the algorithm had obvious advantages over that of the other three algorithms. This guarantees the time efficiency of the decision tree generation and the timeliness of the algorithm with regard to a dramatic increase in air data in the future. From the data in Tables 1 and 2, we can see that the algorithm not only improved the run-time, but also demonstrated relatively high accuracy.

It can be found from the experiments that the improved decision tree prediction model had great breakthroughs in accuracy and computational complexity. Compared with the traditional decision tree algorithm and artificial neural network, we can address a large amount of air quality data more quickly by discretizing the data and introducing weighted coefficients without affecting the accuracy.

In the past, many decision tree algorithms and artificial neural networks paid more attention to the accuracy of the algorithm, but did not pay much attention to the time performance. In the era of big data, the amount of acquired air data is increasing dramatically, and the value of attributes is also increasing. With the increase in nodes, continuous attributes and any attributes in continuous attributes, the efficiency of previous decision tree algorithms will be greatly affected. The time performance will also become an important evaluation index of algorithms for air quality prediction, and the improved decision tree method in this paper provides a new idea for air quality prediction under a large amount of data, which will play an important role in protecting the natural environment and preventing air pollution.

D. AIR DATA MINING IN WUHAN

We used the data from the years 2014-2017 in Wuhan as a training set and the 2018 data as a testing set. In using the historical data of three consecutive years to predict the weather quality in a fourth year, the aim was to test the prediction ability of the approach.

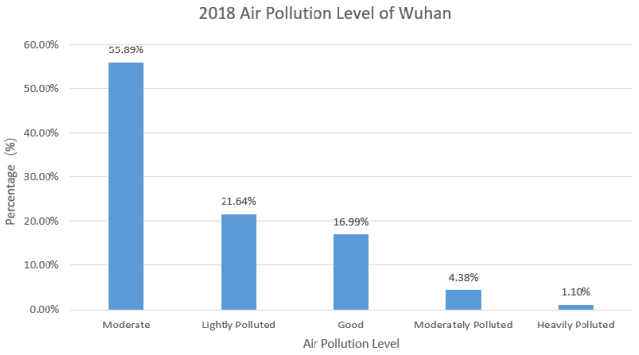

FIGURE 5. Air pollution level of Wuhan.

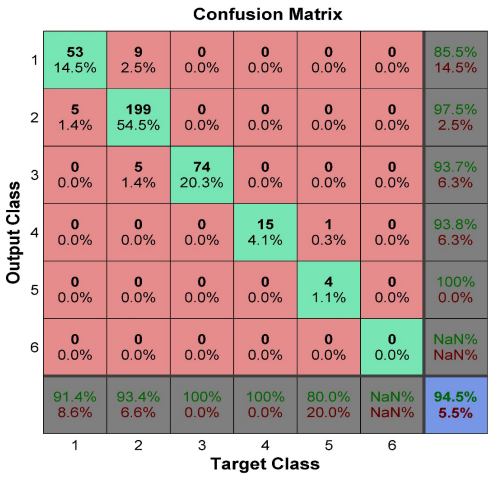

FIGURE 6. Confusion matrix results.

The statistics of the air quality data of Wuhan in 2018 are as shown in Figure 5.

The experimental results are as follows. The confusion matrix diagram is shown in Figure 6.

In Figure 6, each column of the confusion matrix represents a prediction category, and the total number of columns represents the value predicted for that category. Each row represents the true attribution category of the data, and the total number of data instances in each row represents the number of data instances in that category. The diagonal portion displays the number of samples correctly predicted for this category, and the numbers of samples incorrectly predicted for other categories are displayed on both sides. The sum of the diagonal percentages is the accuracy of the model on the test sample set.

The air quality index was divided into six grades: *excellent*, *good*, *light pollution*, *moderate pollution, heavy pollution* and *serious pollution*. There was no serious pollution in Wuhan in 2018. From Figure 6, we can see that 58 samples were predicted in the first category; 53 samples were correctly predicted, and 5 samples belonged to the second category and were misclassified as the first, yielding an accuracy of 91.4%.

There were 213 samples in the second category; 199 samples were correctly predicted, 9 samples in the first category were incorrectly classified into the second category and 5 samples in the third category were incorrectly classified into the second category, yielding an accuracy of 93.4%. There were 74 samples predicted for the third category, with an accuracy of 100%. Fifteen samples were predicted to be in Category 4, with an accuracy of 100%. Four samples were predicted to be in Category 5. One sample belonging to Category 4 was misclassified into Category 5, yielding an accuracy of 80%. There were no sample data in Category 6, so no accuracy of Category 6 could be calculated. The accuracy of each category was as follows: 91.4%, 93.4%, 100%, 100% and 80%. Correspondingly, the error rates were 8.6%, 6.6%, 0%, 0% and 20%. The overall accuracy was 94.5%. It can be seen that the accuracy of the algorithm in predicting the air quality grade was excellent and that moderate pollution needs to be improved.

We used the 2018 air quality data of Wuhan to test the accuracy and compare the run-time with that of four different methods. The accuracy and the run-time data are shown in Table 3.

TABLE 3. Accuracy and run-time comparison on test sets of experiment 3.

Algorithm	Accuracy (%)	Run-time (s)
ID3	90.4%	3.235547
New C4.5	94.5%	1.214771
C4.5	93.9%	2.102696
BP Neural Network	89%	6.170883

It can be seen that the accuracy of the improved C4.5 method was 94.5% from Table 3, which was higher than that of ID3, C4.5 and the BP neural network. Table 3 gives the run-times of the four algorithms. The run-time of the improved algorithm was the least of all the methods. Compared with the neural network model, the run-time of the improved algorithm was obviously less. It is effective and feasible to use the improved C4.5 algorithm to predict future data from historical data.

At the same time, the air quality grade of Wuhan in 2018 was extracted from the experimental results, and the changes over different years were analysed, which provides a good means for improving the algorithm and providing air quality prediction in the future.

VI. CONCLUSION AND FUTURE WORK

The prediction results of the air quality level could provide useful information for safe and reliable societal decisions regarding smart transportation and other public services. However, currently, under a large amount of data, the time performance and accuracy of air quality prediction are urgent problems to be addressed. In this paper, a novel predictive-model-based decision tree method was proposed. The improved model was based on the C4.5 decision tree. The improvements were mainly in two aspects: the feature attribute value and the weighting of the information gain. The accuracy and computational complexity were both improved. Based on the attribute partition principle of the information gain rate, we discretized the data before computing the information entropy. The test attributes in the Fayyad boundary points were determined according to the attributes of the actual data. We selected the attributes with the greatest information gain as the optimal segmentation threshold to replace the information gain rate of traversing all the attributes in the traditional C4.5 algorithm. At the same time, considering the introduction of the weighting coefficient w, the calculation of the gain rate of each attribute was weighted, which reduced the information entropy of some attributes, and correspondingly, improved the information entropy of other attributes. We selected the ROC curve, PR curve and run-time to evaluate the model. The experimental results showed that the improved algorithm significantly improved the time performance and correctness of classification, and it had good prediction ability for future data. When compared with other classification prediction methods, the improved algorithm was slightly superior.

Faced with a large amount of air data, we urgently need an efficient and accurate air quality evaluation algorithm. The first recommended future research work is to compare other classification prediction methods in air quality evaluation at different data levels. Another proposal for future work includes better algorithm performance. The improved algorithm is intended to be used in classification and prediction. Classification based on the improved algorithm is attractive for a large amount of data, which will provide new ideas and methods for future analysis of air quality data, and additional advantages of the improved algorithm can be explored.

REFERENCES

[1] P. Neirotti, A. De Marco, A. C. Cagliano, G. Mangano, and F. Scorrano, "Current trends in Smart City initiatives: Some stylised facts," *Cities*, vol. 38, pp. 25–36, Jun. 2014.

[2] Y. Wang, G. Cao, S. Mao, and R. M. Nelms, "Analysis of solar generation and weather data in smart grid with simultaneous inference of nonlinear time series," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Hong Kong, Apr./May 2015, pp. 600–605.

[3] A. J. Cohen, H. R. Anderson, B. Ostro, K. D. Pandey, M. Krzyzanowski, N, Gutschmidt, K. Pope, A. I. Romieu, J. M. Samet, and K. Smith, "The global burden of disease due to outdoor air pollution," *J. Toxicol. Environ. Health A*, vol. 68, nos. 13–14, pp. 1301–1307, Sep. 2006.

[4] J. Q. Koening. *Health Effects Ofambient Airpollution, How Safe is the Air We Breathe*. Boston, MA, USA: Kluwer, 2000, pp. 1–245.

[5] S. Askari, N. Montazerin, "A high-order multi-variable fuzzy time series forecasting algorithm based on fuzzy clustering," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 2121–2135, Mar. 2015.

[6] V. M. A. Souza, D. F. Silva, and G. E. A. P. A. Batista, "Extracting texture features for time series classification," in *Proc. Int. Assoc Pattern Recognit. (IAPR)*, 2014, pp. 1425–1430.

[7] Y. Sakurai, Y. Matsubara, and C. Faloutsos, "Mining and forecasting of big time-series data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May/Jun. 2015, pp. 919–922.

[8] A. Aggarwal, T. Choudhary, and P. Kumar, "A fuzzy interface system for determining Air Quality Index," in *Proc. Int. Conf. Infocom Technol. Unmanned Syst.*, Dubai, China, Dec. 2017, pp. 786–790.

[9] K. Veljanovska and A. Dimoski, "Air quality index prediction using simple machine learning algorithms," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 7, no. 1, pp. 25–30, Jan./Feb. 2018.
[10] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *Int. J. Environ. Sci. Develop.*, vol. 9, no. 1, pp. 8–16, Jan. 2018.
[11] S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. Abdul, and A. A. Jamal, "Classification model for water quality using machine learning techniques," *Int. J. Softw. Eng. Appl.*, vol. 9, no. 6, pp. 45–52, Jun. 2015.
[12] I. N. Athanasiadis, K. D. Karatzas, and P. A. Mitkas, "Classification techniques for air quality forecasting," in *Proc. 5th ECAI Workshop Binding Environ. Sci. Artif. Intell.*, Riva del Garda, Italy, Aug. 2006, pp. 1–7.
[13] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environ. Sci. Pollut. Res.*, vol. 23, no. 22, pp. 22408–22417, 2016.
[14] Y. Chen, L. Wang, F. Li, B. Du, K.-K. R. Choo, H. Hassan, and W. Qin, "Air quality data clustering using EPLS method," *Inf. Fusion*, vol. 36, pp. 225–232, Dec. 2016.
[15] J. R. Quinlan, "Introduction of decision tree," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.
[16] M. Punia, P. K. Joshi, and M. C. Porwal, "Decision tree classification of land use land cover for Delhi, India using IRS-P6 AWiFS data," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5577–5583, May 2011.
[17] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[18] G. Parthasarathy and B. N. Chatterji, "A class of new KNN methods for low sample problems," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 3, pp. 715–718, May 1990.
[19] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[20] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
[21] [Online]. Available: https://community.tibco.com/wiki/random-forest-template-tibco-spotfirer-wiki-page
[22] A. Graves, *Supervised Sequence Labelling With Recurrent Neural Networks* (Studies in Computational Intelligence), vol. 385. Springer, 2012, pp. 5–13.
[23] Z. C. Lipton, J. Berkowitz, C. Elkan, "A critical review of recurrent neural networks for sequence learning," May 2015, *arXiv:1506.00019*. [Online]. Available: https://arxiv.org/abs/1506.00019
[24] F. A. Gers, "Long short-term memory in recurrent neural networks," M.S. thesis, Dept. Comput. Sci., Swiss Federal Inst. Technol., Lausanne, EPFL, Switzerland, 2001.
[25] W. Bin, "Analyze the spatial-temporal characteristics of air pollution in China by using air pollution index (API)," M.S. thesis, Ocean Univ. of China, Qingdao, China, 2008.
[26] V. M. Niharika and P. S. Rao, "A survey on air quality forecasting techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 103–107, Jan. 2014.
[27] K. Kujaroentavon, S. Kiattisin, A. Leelasantitham, and S. Thammaboosadee, "Air quality classification in Thailand based on decision tree," in *Proc. 7th Biomed. Eng. Int. Conf.*, Fukuoka, Japan, Nov. 2014, pp. 26–28.
[28] K. Fukuda, "Noise reduction approach for decision tree construction: A case study of knowledge discovery on climate and air pollution," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, Mar./Apr. 2007, pp. 697–704.
[29] K. Miloslava and K. Jiří, "Air quality modelling by decision trees in the czech republic locality," in *Proc. 8th WSEAS Int. Conf. Appl. Inform. Commun.*, Rhodes, Greece, Aug. 2008, pp. 196–201.
[30] M. Zhao and X. Li, "An application of spatial decision tree for classification of air pollution index," in *Proc. 19th Int. Conf. Geoinform.*, Shanghai, China, Jun. 2011, pp. 1–6.
[31] Y. Ruiyun, Y. Yang, L. Yang, G. Han, and O. A. Move, "RAQ—A random forest approach for predicting air quality in urban sensing systems," *Sensors*, vol. 16, no. 1, p. 86, Jan. 2016.
[32] X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang, "Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation," *Atmos. Environ.*, vol. 107, pp. 118–128, Apr. 2015.
[33] D. Mishra, "Neuro-fuzzy approach to forecast NO2 pollutants addressed to air quality dispersion model over Delhi, India," *Aerosol Air Qual*, vol. 16, no. 1, pp. 166–174, Jan. 2016.
[34] M. Pawul1 and M. Śliwka, "Application of artificial neural networks for prediction of air pollution levels in environmental monitoring," *J. Ecological Eng.*, vol. 17, no. 4, pp. 190–196, Jan. 2016.
[35] M. Krishan, S. Jha, J. Das, A. Singh, M. K. Goyal, and C. Sekar, "Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India," *Air Qual., Atmos. Health*, vol. 12, no. 8, pp. 899–908, Aug. 2019.
[36] Z. Liangjun, Y. Tan, X. Gang, and X. Shengbing, *Matlab Data Analysis and Data Mining*. Beijing, China: China Mach. Press, 2017, p. 202.
[37] [Online]. Available: https://www.aqistudy.cn/historydata/
[38] J. Langford, "Tutorial on practical prediction theory for classification," *J. Mach. Learn. Res.*, vol. 6, pp. 273–306, Mar. 2005.
[39] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The binormal assumption on precision-recall curves," in *Proc. Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 4263–4266.
[40] J. Kerekes, "Receiver operating characteristic curve confidence intervals and regions," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 2, pp. 251–255, Apr. 2008.
[41] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[42] T. C. W. Landgrebe and R. P. W. Duin, "Efficient Multiclass ROC Approximation by Decomposition via Confusion Matrix Perturbation Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 810–822, May 2008.
[43] L. Wang, Q. Li, Y. Yu, and J. Liu, "Region compatibility based stability assessment for decision trees," *Expert Syst. Appl.*, vol. 105, pp. 112–128, Sep. 2018.
[44] J. Wang, X. Zhang, Z. Guo, and H. Lu, "Developing an early-warning system for air quality prediction and assessment of cities in China," *Expert Syst. Appl.*, vol. 84, pp. 102–116, Oct. 2017.
[45] C. Wu, W. Hu, M. Zhou, S. Li, and Y. Jia, "Data-driven regionalization for analyzing the spatiotemporal characteristics of air quality in China," *Atmos. Environ.*, vol. 203, pp. 172–182, Apr. 2019.
[46] Y. Chen, F. Li, and J. Fan, "Mining association rules in big data with NGEP," *Cluster Comput.*, vol. 18, no. 2, pp. 577–585, Jun. 2015.
[47] T. Lan, Y. Zhang, C. Jiang, G. Yang, and Z. Zhao, "Automatic identification of spread F using decision trees," *J. Atmos. Sol.-Terr. Phys.*, vol. 179, pp. 389–395, Nov. 2018.

YUANNI WANG was born in Zhijiang, Hubei, China, in 1980. She received the B.S. degree in computer science and technology and the M.S. degree in computer applications from the China University of Geosciences, Wuhan, China, in 2003 and 2006, respectively, and the Ph.D. degree from the Wuhan University of China, in 2010. Since 2012, she has been an Assistant Professor with the School of Computer Science, China University of Geosciences. Her research interests include machine learning, intelligent algorithms, and visualization technology.

TAO KONG received the B.S. degree in computer science and technology from the China University of Geosciences, in 2019. His research interests include data mining and high-performance computing.