Research Article

Chao Li, Yimin Shi*, Ping Gao, Yang Shen, Chenchen Ma, and Dawei Shi

Diagnostic model of low visibility events based on C4.5 algorithm

<https://doi.org/10.1515/phys-2020-0007>

Received Nov 27, 2019; accepted Dec 13, 2019

Abstract: In this study the low visibility in Nanjing city is classified and predicted using observed data during 2014 to 2016 with machine-learning based decision tree algorithm (C4.5). For this purpose, the model was trained with 3/4th of the data samples until the self-learning accuracy of the model reached 88.32%. The remaining 1/4th of the data samples were used to verify the model's prediction ability, with the test accuracy reaching 88.34% indicating a good classification diagnosis effect of the model. The results produced with model, generated through learning from the training sample, it is found that the relative humidity, PM10 and PM2.5 are important factors in diagnosing "whether low visibility events will occur in Nanjing": When relative humidity is favorable (*i.e.* <90%) and PM2.5 concentration is not high enough (*i.e.* <146), the probability of low visibility events may reduce; when relative humidity is relatively favorable (*i.e.* ≥ 90%) with a PM10 concentration ≥ 59, low visibility events are more likely to occur; when relative humidity is extremely favorable (*i.e.* ≥ 96%) with a low PM10 concentration (*i.e.* <59), there is also a high probability that low visibility events will occur.

Keywords: Low visibility event; Accuracy; C4.5 algorithm; Diagnostic model

1 Introduction

As an indicator that characterizes atmospheric transparency, visibility is the maximum horizontal distance for the human eye to distinguish the target from the background [1]. Weather phenomena such as fog, haze, precipitation and sand storm, as well as aerosols and polluting gases in the atmosphere, can lead to low visibility events. Having a great impact on people's production and life, low visibility events can cause disasters, especially when it comes to the transportation industry, creating severe visual obstacles for drivers and passengers [2]. Low visibility events may lead to traffic stagnation and can even trigger fatal traffic accidents, bringing about huge economic losses and adverse social impacts [3, 4]. Therefore, there have been many domestic and foreign research studies in this regard. For example, low visibility events frequently occurred in the United States due to man-made pollution during the development of the country [5–8]. Lin [9] analyzed the visibility levels of 4 major cities in China, *i.e.* Beijing, Shanghai, Guangzhou and Chengdu. Di *et al.* [10] studied the trends of visibility levels in 6 major cities in China, *i.e.* Beijing, Chengdu, Guangzhou, Shanghai, Shenyang and Xi'an, from 1973 to 2007. Fan *et al.* [11] analyzed the changes in atmospheric visibility caused by human-induced pollution in the Beijing-Tianjin-Hebei Region by removing the influence of meteorological factors on visibility. Li *et al.* [12] and Liu *et al.* [13] analyzed the physical structures of different heavy fog processes in Nanjing. Wu *et al.* [14] studied the macroscopic and microscopic structure of the heavy fog at the Expressway in the Nanling Dayaoshan Mountain. According to Zhang [15, 16], the chemical components of atmospheric aerosol not only causes air pollution, but also affects the climate. In recent years, low visibility events have occurred more frequently in economically developed areas of China such as the Yangtze River Delta, the Pearl River Delta, and the North China Plain [17]. With the rapid industrial development and the acceleration of urbanization, air pollution is becoming viable threat to human health and atmosphere, which makes the causes of low visibility events even more complicated. The increasing intensity density of air pollu-

***Corresponding Author: Yimin Shi:** Lianyungang Meteorological Bureau, Lianyungang 222006, China; Email: 395158610@qq.com

Chao Li, Yang Shen: Jiangsu Meteorological Station, Nanjing 210008, China; Key Laboratory of Transportation Meteorology, China Meteorological Administration, Nanjing 210009, China

Ping Gao: Jiangsu Meteorological Station, Nanjing 210008, China; Jiangsu Meteorological Public Service Center, Nanjing 210008, China

Chenchen Ma, Dawei Shi: Laboratory of Transportation Meteorology, China Meteorological Administration, Nanjing 210009, China; Lianyungang Meteorological Bureau, Lianyungang 222006, China

tion not only affect the weather and climate through radiative forcing and altering the energy and water balance but also have many health-related consequences. Hence, studying the low visibility events based on meteorological data and atmospheric composition data, we can improve our understanding of the underlying dynamics and mechanism of the low visibility events.

Statistical and numerical computation techniques and their application in atmospheric sciences always provide significant information in the understanding and unravelling the mechanism, drivers and associated dynamics [18–24]. With the continuous improvement in the computer performance, data mining technology has been increasingly used in scientific research especially in the field of meteorological and environmental sciences. For instance, Zhang *et al.* [25, 26] used the C4.5 algorithm to predict the typhoon track and trajectory in the North Pacific with accuracy of more than 80%. Geng *et al.* (2016) used the finite mixture model (FMM) algorithm and the CART decision tree algorithm to classify the path and predict the frequency of tropical cyclones landing in China, and the prediction accuracy for the frequency of re-curvature typhoon was more than 85% [27]. David *et al.* [28] used the random forest (RF) algorithm to establish a mesoscale convective system (MCS) based on radar data, satellite data and model output data.

In the current study, the C4.5 decision tree algorithm is applied to explore and establish the classification of low visibility events in Nanjing using meteorological and atmospheric composition data. Considering the accuracy and low complexity of the decision tree algorithm [29], it is expected that this approach will provide new insights and methods for the diagnosis of low visibility events in other regions, thereby offering a scientific reference for the assessment and prevention of low visibility events.

2 Materials and methods

2.1 Distribution of stations and source of data

Nanjing is the capital city of Jiangsu province, which is located in southwest corner of Jiangsu Province, an important core city in the Yangtze River Delta urban agglomeration, the most economically developed region in China. The population density of Nanjing is more than 8 million and a total area of 6,587 km², the urban population of Nanjing is 678,140,000, with an urbanization rate of 82%. Nanjing is one of the most important megacities in the

Yangtze River Delta as well as in the East China. With Nanjing as the representative city of the Yangtze River Delta, the prediction methods of low visibility were tested as pilot study in the city. In order to reduce the uncertainties associated with data quality due to the distance between the environmental monitoring stations and the meteorological stations, The atmospheric composition data used was obtained from the environmental monitoring station on Shanxi Road and the meteorological data from the Nanjing Meteorological Station. As shown in Figure 1, the distance between the environmental monitoring station and the meteorological station is less than 20km.

The hourly surface and near surface meteorological observations (temperature, air pressure and relative humidity at a height of 2m, surface temperature at a height of 0cm, visibility, wind speed at a height of 10m, wind direction at a height of 10m, and precipitation) of Nanjing city for the duration of 2014–2016 were collected from 58038 Meteorological Observation Station through Jiangsu Provincial Meteorological Bureau. Compared with previous manual observation data, this set of data features higher observation frequency and no subjective observation error in terms of visibility. To make the distance between the meteorological observation station and the environmental monitoring station as small as possible and the data complete and matched, the atmospheric composition data observed on hourly scale from 2014 to 2016 was used in this study. The atmospheric composition data include fine particulate matter (PM_{2.5}), inhalable particulate matter (PM₁₀), ozone (O₃), sulfur dioxide (SO₂), nitrogen dioxide (NO₂) and carbon monoxide (CO). The meteorological data and atmospheric composition data were integrated into a data set, and the missing data was eliminated. There are a total of 22,658 pieces of data, accounting for 86% of the data collected during the total time (201,430h) from 2014 to 2016, so the data is representative.

2.2 C4.5 decision tree algorithm

The C4.5 decision tree algorithm, a classification and prediction algorithm was invented in 1979 by JR Quinlan who also proposed the ID3 algorithm for discrete attribute data. Then the ID3 algorithm was continuously improved to produce the C4.5 algorithm, which was added with discretization of continuous attributes [30].

This algorithm selects the classification attribute on each node based on the size of the attributed information gain, split the samples according to the attribute that can bring the maximum information gain, and recursively split the samples until the stop condition is reached [25–27]. In

Figure 1: Geographical location of Nanjing, the schematic map of the location of meteorological observation station and environmental monitoring station (▲ represents the location of environmental monitoring station, and ■ represents the location of meteorological observation station).

the end, the conclusion are tested to trim and reject the subsets of samples without making significant contributions to the model.

Let S be the training set including s data samples, $S(C_i)$ be the number of samples belonging to the class C_i ($i = 1, 2, \dots, m$) in S , then the probability that samples in the training set belong to the i -th class can be expressed as follows:

$$P(C_i) = \frac{S(C_i)}{s} \quad (1)$$

The entropy of the training set S is defined as follows:

$$\begin{aligned} info(S) &= - \sum_{i=1}^m P(C_i) \log_2 P(C_i) \\ &= - \sum_{i=1}^m \frac{S(C_i)}{s} \log_2 \left[\frac{S(C_i)}{s} \right] \end{aligned} \quad (2)$$

Then, information S can be divided into $\{S_1, S_2, \dots, S_v\}$ based on attribute A , then the information entropy of the leaf node for the classification information is as follows:

$$info(A|S) = - \sum_{j=1}^v \frac{S_j}{S} info(S) \quad (3)$$

Then the information gain can be calculated as]:

$$gain(A|S) = info(S) - info(A|S) \quad (4)$$

The gain rate is:

$$gainratio(A|S) = \frac{gain(A|S)}{info(A|S)} \quad (5)$$

The C4.5 algorithm is developed based on the ID3 algorithm. One of the major improvements is that the C4.5 algorithm can process continuous data. The C4.5 algorithm

Table 1: Binary target variable samples in the training sets and test sets balanced by playback sampling.

	Number of samples representing non-low visibility events	Number of samples representing low visibility events	Number of samples representing low visibility events after sampling with replacement
Training set	16451	543	16433
Test set	5473	191	5472

Figure 2: Strategic schematic for selecting training samples.

generally processes continuous attribute data as follows: 1) Sort the attribute data among the nodes; 2) Dynamically divide data in the training set with different thresholds; 3) With the midpoint of the two values at both ends of the input data as the threshold, determine the new threshold as the input changes; 4) Determine the 2 categories according to the threshold, and divide all the data samples into the 2 categories; 5) Obtain all possible thresholds and calculate the information gain and gain rate under different conditions; 6) In the end, each continuous attribute will be divided into 2 categories by the threshold (greater than or equal to the threshold and less than the threshold).

Cross validation was adopted to test the effect of the algorithm model. In other words, part of the data was used to train the model, that is, the training set. Another part of the independent data was used to test the model, that is, the test set [28–30]. The sample size of the training set is usually about 3 times that of the test set. The self-learning accuracy refers to the ratio of the number of samples correctly trained to the total number of samples trained, and the test accuracy is the ratio of the number of samples correctly tested to the total number of samples tested.

3 Classification diagnosis model of low visibility events in Nanjing based on C4.5 algorithm

3.1 Preprocessing of experimental data

In this study, the events with atmospheric visibility < 1 km was defined as the low visibility events. The C4.5 algorithm, a supervised algorithm in data mining, was used to establish a classification diagnosis model of the low visibility events (visibility < 1 km) in Nanjing. Firstly, “the occurrence of the low visibility events in Nanjing” was abstracted into a binary classification problem *i.e.*, when visibility < 1 km, the low visibility event will occur; while when visibility ≥ 1 km, the low visibility event will not occur. Among the 22,658 data samples, the low visibility events occurred to 734 data samples, while the non-low visibility events happened to 21,924 data samples in Nanjing during the study period. In order to minimize the influence on distribution characteristics of the target data, the samples in the test set were selected on the basis of the equidistant sampling *i.e.*, with time as the order, the 4th–4n-th data samples were selected into the test set, and the remaining data samples were classified into the training set (Figure 2).

The training set and the test set were counted separately (Table 1). The training set consists of 16,994 samples, out of which 543 samples represent the low visibility events, while 16,451 samples stand for non-low visibility events. Similarly, the test set contains 5,664 samples, out of which 191 samples signify the low visibility events, and 5,473 samples represent the non-low visibility events. Due to the objective weather factors, the frequency of low visibility events is significantly lower than that of non-low visibility events. Therefore, there is a huge difference between the sample size of low visibility events and non-low visibility events in the training set and the test set. If the model had been built directly based on the data sets, it would have been difficult to objectively reflect the prediction results. To make the model results more objective,

Table 2: Diagnostic Rule Set of Meteorological and Atmospheric Component Observation Elements for Low Visibility Events in Nanjing Discovered by the C4.5 Algorithm.

Decision rule	Decision attributes	Learning accuracy
A : If (relative humidity < 90% and PM _{2.5} < 146 mg/m ³), then low visibility events will not happen	Relative humidity and PM _{2.5}	12232/12324=99.25%
B : If (relative humidity < 90% and PM _{2.5} > 146 mg/m ³), then low visibility events will happen	Relative humidity and PM _{2.5}	879/1621=54.23%
C : If (relative humidity ≥ 90% and PM ₁₀ ≥ 59 mg/m ³), then low visibility events will happen	Relative humidity and PM ₁₀	13070/14899=87.72%
D : If (relative humidity ≥ 90% and PM ₁₀ < 59 mg/m ³ and relatively humidity < 96 mg/m ³), then low visibility events will not happen	Relative humidity and PM ₁₀	1231/1991=61.83%
E : If (relative humidity ≥ 96% and PM ₁₀ < 59 mg/m ³), then low visibility events will happen	Relative humidity and PM ₁₀	1632/2049=79.65%

a random repeated sampling with replacement was conducted on the target samples, representing low visibility events in the training and the test sets. The number of target samples aims to maintain the low visibility events and non-low visibility events in the training and test sets in the same order of magnitude.

3.2 Low visibility event classification diagnosis model based on C4.5 algorithm and statistical analysis

With “whether it is a low visibility event” as the target variable of the model, a number of meteorological elements such as temperature, air pressure, relative humidity (RH), average wind speed at 10 min, average wind direction at 10 min and surface temperature at a height of 0 cm as well as atmospheric composition data such as PM₁₀, PM_{2.5}, O₃, SO₂, NO₂ and CO as the input variables of the model, the pre-processed training set was calculated on the basis of the C4.5 algorithm to obtain the decision tree (Figure 3).

The decision tree is an inverted tree diagram, where the root node *i.e.*, the top node of the decision tree, denotes relative humidity, which is the most heterogeneous attribute worked out by the algorithm based on the information gain. Moreover, the PM_{2.5} and PM₁₀ are located on the second layer of the decision tree and are less important than relative humidity; however, these variables are still important attributes to determine the occurrence of the low visibility events. Similarly, some data attributes are not found in the decision tree, indicating that such data are not heterogeneous enough and do not play an important role in assessing the occurrence of the low visibility events.

Figure 3: Diagnostic Decision Tree Model of Meteorological and Atmospheric Component Observation Elements for Low Visibility Events in Nanjing Area from 2014 to 2016.

A rule set for analyzing the occurrence of the low visibility events in Nanjing using the decision tree. The model based on the training set has shown an overall self-learning accuracy of 88.32%, and each rule has its own self-learning accuracy relative to their actual situations. The pre-processed test set was then used to test the generalization ability of the model, with the test accuracy being 88.34%. After a thorough verification, it is found that the model has a good overall classification effect and strong generalization ability, providing a concise, understandable, and valuable reference for diagnosing the occurrence of low visibility events.

As shown in Table 2, Rules A, C and E have relatively high learning accuracy and apply to a large proportion of samples. However, although Rules B and D occupy a relatively small proportion of samples, the learning effect is relatively poor. Based on a statistical analysis of the data of Rules B and D in different situations (with or without low visibility events), Table 3 was obtained.

Table 3: Statistics of Some Meteorological and Environmental Elements in Rule B and Rule D.

Data Category	Average relative humidity (%)	Average air temperature (°C)	Average precipitation (mm)	Ground-air temperature difference (°C)	Average visibility (km)
Rule B	69.24	12.40	0.01	0.57	2.58
Rule D	92.92	17.53	0.70	0.35	4.57
Rule B Non-low visibility	68.64	12.57	0.01	0.64	2.65
Rule B Low visibility	85.57	7.99	0.00	−1.20	0.79
Rule D Non-low visibility	92.90	17.52	0.63	0.35	4.65
Rule D Low visibility	94.10	18.11	4.18	−0.03	0.75

According to Table 3, for Rule B, low visibility occurs when the temperature is significantly low and no precipitation occurs, and there should be certain temperature inversion *i.e.*, the ground-temperature difference (0 cm ground temperature-2 m temperature) should be smaller than 0. For Rule D, it is easier to diagnose errors when there is strong precipitation and ground-temperature difference is smaller than 0.

Based on Figure 3 as well as Tables 2 and 3, the following conclusions are presented: (1) Relative humidity is the most important attribute to determine the occurrence of the low visibility events; (2) Rule A in Table 2 indicates that when relative humidity is unfavorable (*i.e.* <90%) and PM_{2.5} concentration is not high enough (*i.e.* <146mg/m³), the low visibility events are generally unlikely to happen; (3) According to Rule C in Table 2, when relative humidity is favorable (*i.e.* ≥ 90%) and the PM₁₀ concentration is ≥ 59mg/m³, the low visibility events are more likely to occur; (4) when relative humidity is extremely favorable (*i.e.* ≥ 96%) and the PM₁₀ concentration is <59mg/m³, there is also a high probability of occurrence of the low visibility events; (5) Many meteorological elements and pollutant gases are not reflected in the Nanjing low visibility event prediction model. From the perspective of data heterogeneity, relative humidity and PM concentration (PM_{2.5} and PM₁₀) are more important than other meteorological elements and polluting gases for assessing the occurrence of the low visibility events; (6) A combination of the objective decision tree model and subjective statistical analysis can improve the diagnosis accuracy of the low visibility events.

4 Conclusion

In this study, we defined the low visibility event and analyzed its statistical characteristics in Nanjing, China. Based on observed hourly data of the meteorological and

atmospheric composition for the period 2014 to 2016, the classic C4.5 decision tree algorithm in data mining was used to establish the Nanjing low visibility classification and prediction model, which achieved a good prediction effect. The proposed model suggested the following conclusions:

1. The classic C4.5 decision tree algorithm in data mining was used to establish the Nanjing low visibility classification and prediction model based on the meteorological and atmospheric composition data hourly observed from 2014 to 2016 in Nanjing. The model has performed well with a self-learning accuracy of 88.32% and a test accuracy of 88.34% (2) The occurrence of the low visibility events is closely related to relative humidity and the concentration of PM_{2.5} and PM₁₀, where low visibility events may occur under different conditions of relative humidity and PM concentration. (3) The low visibility events were divided into 2 types (Rules B and D) by the decision tree model. It has been found statistically that low temperature, temperature inversion and precipitation all can contribute to the occurrence of low visibility events in Nanjing. With the arrival of the big data era and the continuous advancement of artificial intelligence, it becomes an inevitable development trend to explore the frequent occurrence of natural disasters from the perspectives of machine learning and data mining. In addition, the continuous research on the disaster mechanism will also facilitate a better application of machine learning and data mining in the field of disaster management.

Acknowledgement: This study is supported by the following funds: Youth Fund Project of Jiangsu Provincial Meteorological Bureau (KQ201802), Huaihe River Basin Meteorological Open Research Fund (HRM201602), “Marine Weather Forecast Technology” Innovation Team of Jiangsu

Provincial Meteorological Bureau, Project of Department of Science & Technology Jiangsu Province (BE2011720), and Science and Technology Support Project of Lianyungang City (SH1634).

References

- [1] Sheng PX, Mao JT, Li JG, *et al.* Atmospheric physics. Beijing: Peking University Press; 2003.
- [2] Li ZH, Liu DY, Yang J, *et al.* Physical and chemical characteristics of winter fogs in Nanjing. *Acta Meteorol Sin.* 2011;69(4):706–18.
- [3] Gultepe I, Tardif R, Michaelides SC, Cermak J, Bott A, Bendix J, *et al.* Fog research: a review of past achievements and future perspectives. *Pure Appl Geophys.* 2007;164(6/7):1121–59.
- [4] Black AW, Villarini G, Mote TL. Effects of Rainfall on Vehicle Crashes in Six U.S. States. *Weather Clim Soc.* 2017;9(1):53–70.
- [5] Malm WC, Sisler JF, Huffman D, Eldred RA, Cahill TA. Spatial and seasonal trends in particle concentration and optical extinction in the United States. *J Geophys Res.* 1994;99 D1:1347–70.
- [6] Malm WC, Sisler JF, Pitchford ML, *et al.* IMPROVE (Interagency Monitoring of Protected Visual Environments): Spatial and seasonal patterns and temporal variability of haze and its constituents in the United States: Report III. CIRA Report. Fort Collins: CIRA; 2000.
- [7] Trijonis J. Visibility in California. *J Air Pollut Control Assoc.* 1982;32(2):165–9.
- [8] Trijonis J, Shapland D. Existing Visibility Levels in the U.S., Iso-pleth Maps of Visibility in Suburban/Nonurban Areas During 1974–1976. EPA-450/5-79-101. U.S. Environmental Protection Agency; 1979.
- [9] Lin M, Tao J, Chan CY, Cao JJ, Zhang ZS, Zhu LH, *et al.*; Agency Lin M. Regression analyses between recent air quality and visibility changes in megacities at four haze regions in China. *Aerosol Air Qual Res.* 2012;12(6):1049–61.
- [10] Chang D, Song Y, Liu B. Visibility trends in six megacities in China 1973–2007. *Atmos Res.* 2009;94(2):161–7.
- [11] Fan YQ, Li CQ. Study on the trend of atmospheric visibility change in Beijing. Tianjin and Hebei from 1980 to 2003. *Plateau Meteorology* 27(6).
- [12] Li ZH, Huang JP, Zhou YQ, *et al.* Physical Structures of the Five-Day Sustained Fog around Nanjing in 1996. *Acta Meteorol Sin.* 1999;57(5):622–31.
- [13] Li ZH. Studies of Fog in China over the Past 40 Years. *Acta Meteorol Sin.* 2001;59(5):616–24.
- [14] Wu D, Deng XJ, Mao JT, *et al.* A Study on Macro-and Micro-Structures of Heavy Fog and Visibility at Freeway in the Nanling Dayaoshan Mountain. *Acta Meteorol Sin.* 2007;65(3):406–15.
- [15] Zhang XY. Aerosol over China and Their Climate Effects. *Diqu Kexue Jinzhan.* 2007;22(1):12–26.
- [16] Zhang XY. Characteristics of the chemical components of aerosol particles in the various regions over China. *Acta Meteorol Sin.* 2014;72(6):1108–17.
- [17] Chen J, Zhao CS. A Review of Influence Factors and Calculation of Atmospheric Low Visibility. *Advances in Met S&T.* 2014(4), 44–51.
- [18] Zhou JB, Huang JY. Recent advances in statistical meteorology in China. *Acta Meteorol Sin.* 1997;55(3):297–305.
- [19] Gao W, Wang W. A Tight Neighborhood Union Condition on Fractional (G, F, N^*, M) -Critical Deleted Graphs. *Colloq Math-Warsaw.* 2017;149(2):291–8.
- [20] Gao W, Wang W. New Isolated Toughness Condition for Fractional (G, F, N) -Critical Graph. *Colloq Math-Warsaw.* 2017;147(1):55–65.
- [21] Baig AQ, Naeem M, Gao W. Revan and hyper-Revan Indices of Octahedral and Icosahedral Networks. *Applied Mathematics & Nonlinear Sciences.* 2018;3(1):33–40.
- [22] Dewasurendra M, Vajravelu K. On the Method of Inverse Mapping for Solutions of Coupled Systems of Nonlinear Differential Equations Arising in Nanofluid Flow, Heat and Mass Transfer. *Applied Mathematics & Nonlinear Sciences.* 2018;3(1):1–14.
- [23] Khellat F, Khormizi MB. A Global Solution for a Reaction-Diffusion Equation on Bounded Domains. *Applied Mathematics & Nonlinear Sciences.* 2018;3(1):15–22.
- [24] Lakshminarayana G, Vajravelu K, Sucharitha G, Sreenadh S. Peristaltic Slip Flow of a Bingham Fluid in an Inclined Porous Conduit with Joule Heating. *Applied Mathematics & Nonlinear Sciences.* 2018;3(1):41–54.
- [25] Zhang W, Leung Y, Chan JC. The Analysis of Tropical Cyclone Tracks in the Western North Pacific through Data Mining. Part I: Tropical Cyclone Recurvature. *J Appl Meteorol Climatol.* 2013;52(6):1394–416.
- [26] Zhang W, Leung Y, Chan JC. The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part II: tropical cyclone landfall. *J Appl Meteorol Climatol.* 2013;52(6):1417–32.
- [27] Geng H, Shi D, Zhang W, Huang C. A prediction scheme for the frequency of summer tropical cyclone landfalling over China based on data mining methods. *Meteorol Appl.* 2016;23(4):587–93.
- [28] David A, James OP, John KW, *et al.* Probabilistic Forecasts of Mesoscale Convective System Initiation Using the Random Forest Data Mining Technique. *Wea Forecasting.* 2016;31(2):581–99.
- [29] Shi D, Li C, Shi Y, *et al.*; Study on the Localization Diagnosis of Extra Heavy Fog on the Background of the Fog Weather Based on Machine Learning Algorithms. *Zaihaixue.* 2018;33(2):193–9.
- [30] Quinlan J. Decision trees as probabilistic classifiers. *Proc. Fourth Int. Workshop on Machine Learning, Irvine, CA, American Association for Artificial Intelligence, 1987.* <https://doi.org/10.1016/B978-0-934613-41-5.50007-6>.