J Inf Process Syst, Vol.20, No.6, pp.718~730, December 2024
https://doi.org/10.3745/JIPS.04.0324
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Application Research of Rainfall Prediction Based on Optimized Machine Learning Algorithm in Meteorological Data

Daoqing Gong, Cheng Yuan, Xinyan Gan*, Xiang Gao, and Guizhi Sun

Abstract

In recent years, the rapid development of artificial intelligence technology has brought new opportunities to the meteorological field. Specifically, machine learning (ML) algorithms have proven valuable tools in rainfall retrievals, demonstrating the practicability of using ML algorithms when facing high-dimensional and complex data. By collecting data and using ML algorithms to mine and analyze the data, ML models can solve the problem of rainfall prediction in meteorology. Spurred by this advantage, this paper compared five ML algorithms for rainfall prediction using the National Population Health Science data from China, and the five ML algorithms were optimized appropriately. The data employed was first preprocessed to find and fill in the missing values, remove duplicate values, mine the correlation between data features, and generate visual results. Then, logistic regression, k-nearest neighbor algorithm, naive Bayes, decision tree algorithms, and random forest were used to mine and analyze the meteorological data for weather prediction. Finally, the performance of the models before and after optimization is compared to provide decision support for rainfall prediction.

Keywords

Machine Learning, Data Mining, Optimization Model, Meteorological Data, Rainfall Forecast

1. Introduction

Meteorology is closely related to human beings. Compared to nature, humans are extremely insignificant, various adverse weather conditions can pose a significant threat to human life and health. Analyzing and forecasting extreme weather using observations from satellites and weather stations is a daunting task. Accurate weather forecasts are essential for ensuring a good quality of life for people. Unlike other fields, meteorology possesses vast historical observational datasets, offering immense potential for the application of machine learning. However, the potential for false detections and alarms in actual data necessitates extensive data cleaning efforts. In recent years, the application of big data and machine learning methods in the field of meteorology has become increasingly popular. This paper leverages an optimized machine learning algorithm to analyze meteorological data from the National Population Health Science Data Center, conducting in-depth research on rainfall prediction.

Manuscript received July 5, 2022; first revision October 14, 2022; second revision January 16, 2023; accepted January 24, 2023.
***Corresponding Author:** Xinyan Gan (249743155@qq.com)
School of Public Health and Management, Guangxi University of Chinese Medicine Nanning, 530200, China (605445592@qq.com, 906268957@qq.com, 249743155@qq.com, 371590205@qq.com, 80978984@qq.com)
Daoqing Gong and Cheng Yuan contributed equally to this work and were recognized as co-first authors.

With the integration of artificial intelligence into meteorology, researchers can leverage machine learning algorithms to predict evolving weather patterns. These algorithms can facilitate the mining and analysis of historical data, enabling the extraction of pertinent information from vast meteorological datasets, thereby enhancing the accuracy of future weather predictions [1]. Data mining techniques can uncover previously unknown and potentially valuable insights from extensive databases, offering advanced technical support for decision-making and analysis [2,3].

In recent years, research on meteorological prediction using data mining technology has primarily focused on two approaches: the mathematical statistical methods based on statistics, and the methods utilizing machine learning and soft computing [4]. In the study by Alrashidi and Qamar [5], factory load and meteorological data were recorded hourly from 2016 to 2017. After applying data pre-processing techniques, various machine learning algorithms were implemented and compared to predict factory load.

Due to the inherent complexity and uncertainty of meteorological data, demonstrate limited accuracy in weather prediction. Data mining technology enable the discovery of the internal relationship between various meteorological parameters in historical meteorological data and atmospheric dynamics, and get various underlying patterns to reveal the future weather changes. This methodology has demonstrated significant implications in meteorological research and applications. Common machine learning algorithms employed in meteorological studies include logistic regression (LR), k-nearest neighbor (KNN), naive Bayes (NB), decision tree (DT), and random forest (RF).

However, these machine learning algorithms such as LR, KNN, NB, DT, and RF each exhibit certain limitations. For example, the LR algorithm is susceptible to underfitting and its classification accuracy is too low. The DT algorithm struggles with handling the missing data and tends to ignore the correlation among attributes in the data set, which potentially leading to overfitting in the training model. When the RF algorithm includes a large number of decision trees, the space and time required for training increase significantly. There are many aspects of the random forest that are difficult to explain, as it is somewhat of a black box model. In some noisy sample sets, the model is easy to fall into overfitting. The k value of KNN algorithm presents a challenge, and the choice of category must consider its appropriateness; how to choose the appropriate distance measurement is an existing problem. The NB algorithm is sensitive to the preparation and quality of the input data. To solve the above problems, this paper optimizes five traditional machine learning algorithms from different perspectives. The experimental results show that the optimized model has achieve significant improvements.

This paper is organized as follows: Section 2 normalizes the data and analyzes the correlation. In Section 3, we use various machine learning and optimized models to predict precipitation. In Section 4, we analyze and compare the prediction results of various algorithms in terms of accuracy and error. Section 5 concludes this paper.

2. Data Analysis and Processing

The meteorological data in this paper were obtained from the data warehouse of China National Population Health Science Data Center [6]. The data used in this paper includes a total of 34 meteorological stations, and 130,203 records from 2008 to 2018 were collected. In 2018, with only half a year of data, there are 15 features.

2.1 Data Analysis

In this data analysis section, a specific station is selected as a case study, with the station number being 50,953 and containing 3,831 pieces of data.

Fig. 1 illustrates the relationship between rainfall and rainfall frequency of different features of the original data. As can be seen from this figure, rainfall less than 0 indicates drought. The proportion with rainfall of 0 is high because it is affected by all the features. There is not a single instance with a rainfall amount of 0.5. The proportion of rainfall amounting to 1 is relatively small, predominantly influenced by wind speed and pressure.

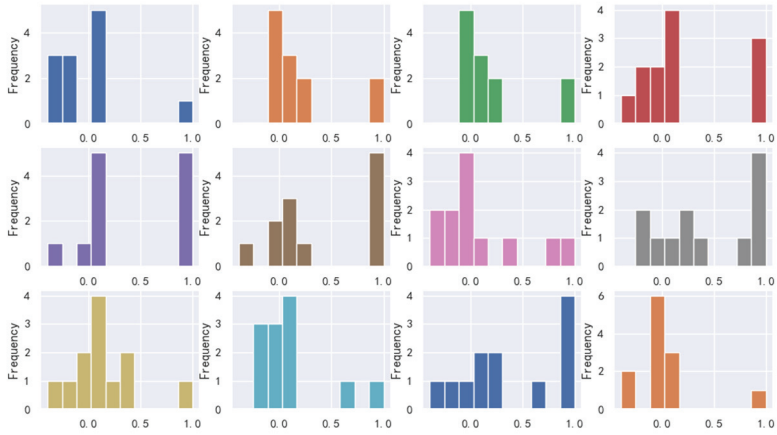
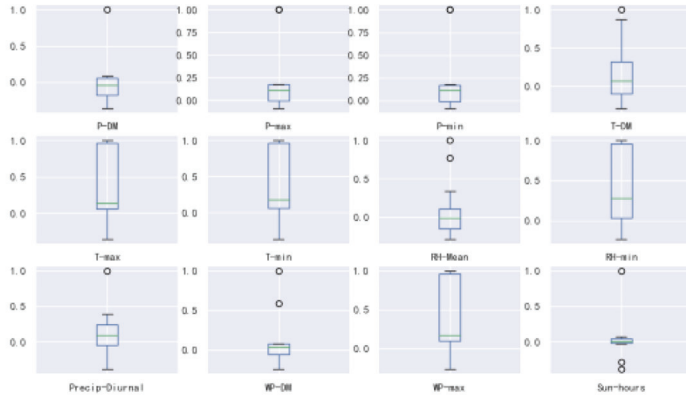


Fig. 1. Relationship between rainfall and rainfall frequency.

	Station	P-DM	...	WP-max	Sun-hours
count	3831.0	3831.000000	...	3831.000000	3831.000000
mean	50953.0	10174.738449	...	59.882015	67.342469
std	0.0	2075.692179	...	528.996535	529.886462
min	50953.0	9671.000000	...	13.000000	0.000000
25%	50953.0	9915.000000	...	37.000000	24.000000
50%	50953.0	9978.000000	...	48.000000	63.000000
75%	50953.0	10054.000000	...	61.500000	90.000000
max	50953.0	32766.000000	...	32766.000000	32766.000000

(a)



(b)

Fig. 2. (a) Descriptive statistics and (b) box-plot.

The descriptive statistical analysis presented in Fig. 2(a) illustrates the overall distribution of rainfall data and the correlations between rainfall characteristics. The central tendency and dispersion degree of rainfall data can be observed through statistical values. The right side of Fig. 2(b) shows a box-plot. A box-plot is a statistical plot used to show the dispersion of a set of data. Box-plot is a method to describe data by using five statistics in the data: minimum value, first quartile (25th percentile), median, third quartile (75th percentile), and maximum value. It also provides insight into the data's symmetry, spread, and other characteristics, especially for the comparison of several samples. As shown in Fig. 2, the original data has many values that need to be processed urgently, so the next step of data preprocessing is required.

2.2 Data Pre-processing based on Pandas

Pandas is widely used for data cleaning, mining, analysis and prediction. sklearn refers to the machine learning library scikit-learn encapsulated by Python, which can be used in all aspects from data pre-processing to model training [7]. The sklearn implementation utilized in this study include LR, KNN, NB, DT, and RF.

Data normalization: The original data [6] used in this paper is not suitable for direct calculation. Therefore, this paper uses Panda library to clean the original data. First, the data shall be normalized, and all characteristic values of meteorological data are scaled to the (0,1) interval. The normalization results are shown in Fig. 3.

	P-DM	P-max	P-min	T-DM	T-max	T-min	RH-Mean	RH-min	WP-DM	WP-max	Sun-hours
0	0.89926	0.15161	0.89697	0.20424	0.18343	0.19254	0.56250	0.49000	0.15833	0.06729	0.41892
1	0.89802	0.15144	0.90387	0.20877	0.20266	0.22239	0.69792	0.62000	0.13333	0.05477	0.40541
2	0.89926	0.15274	0.89721	0.27080	0.25592	0.28955	0.68750	0.52000	0.11667	0.04695	0.25000
3	0.90642	0.15303	0.90707	0.21634	0.21154	0.22836	0.72917	0.63000	0.10833	0.04851	0.38514
4	0.88642	0.15090	0.88957	0.27383	0.26923	0.27313	0.70833	0.64000	0.16667	0.07512	0.24324
5	0.89778	0.15198	0.89524	0.23601	0.21006	0.25970	0.63542	0.38000	0.15000	0.07355	0.43919
6	0.89037	0.15173	0.89081	0.22390	0.21450	0.23731	0.59375	0.39000	0.19167	0.06260	0.41892
7	0.89086	0.15048	0.89475	0.20726	0.19379	0.23433	0.47917	0.34000	0.12500	0.07668	0.48649
8	0.90667	0.15299	0.90436	0.15129	0.14941	0.18060	0.44792	0.34000	0.17500	0.07825	0.47297
9	0.90494	0.15307	0.90732	0.18608	0.18343	0.20448	0.55208	0.45000	0.12500	0.06103	0.46622
10	0.91827	0.15541	0.91915	0.16793	0.20266	0.16716	0.63542	0.43000	0.15833	0.06886	0.39865
11	0.93901	0.15859	0.93320	0.10287	0.07692	0.12537	0.56250	0.38000	0.12500	0.05477	0.27027
12	0.94716	0.15942	0.95218	0.10893	0.09024	0.13433	0.51042	0.35000	0.20000	0.06416	0.44595
13	0.93901	0.15884	0.94405	0.08775	0.09911	0.10149	0.55208	0.38000	0.13333	0.05634	0.40541
14	0.93877	0.15846	0.94084	0.10893	0.09024	0.12239	0.56250	0.45000	0.14167	0.05477	0.46622
15	0.94247	0.15875	0.94774	0.10590	0.12130	0.09851	0.56250	0.41000	0.14167	0.05790	0.23649
16	0.94222	0.15888	0.94701	0.11498	0.12870	0.11343	0.58333	0.48000	0.11667	0.05164	0.41892

Fig. 3. Data normalization.

Data duplicate value processing: There are some duplicate values and missing values in the original data. For the duplicate values, the deletion method is used in this paper, and only one value is retained; for missing values, this paper uses the mean filling method. Fig. 4 shows the statistical results before data processing.

Data feature correlation analysis: Correlation analysis involves examining the relationship between two or more variables to assess the degree of association between them. In this paper, the correlation analysis of meteorological data features is carried out, and the analysis results are shown in Fig. 5.

As observed in Fig. 5 the light-yellow grid represents strong correlation, for example, the correlation between T-Max and Rh min, as well as T-max and WP-max is 0.96; however, a negative correlation exists between some features; for example, the correlation between Sun-hours and Diurnal is -0.39.

```
-------------Statistical results before data processing-----------
P-DM: 32
P-max: 33
P-min: 33
T-DM: 2
T-max: 1
T-min: 1
RH-Mean: 5
RH-min: 1
Precip-Diurnal: 1
WP-DM: 2
WP-max: 1
Sun-hours: 1
```

Fig. 4. Statistical results before data processing.

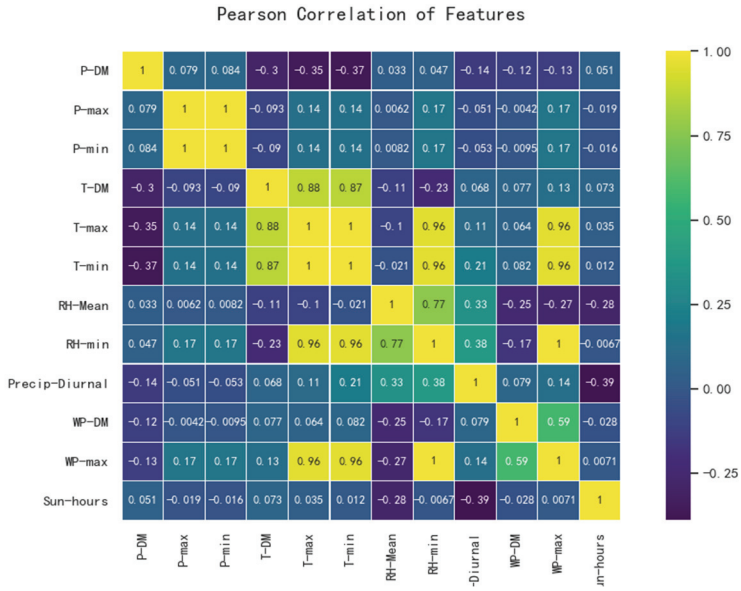


Fig. 5. Data feature correlation analysis.

3. Based on Machine Learning and Its Optimization Algorithm

In recent years, researchers have increasingly applied machine learning algorithms and data mining methods to meteorological forecasting, aiming to enhance the understanding of meteorological laws and the ability of weather forecasting. This approach has garnered significant attention from experts and scholars in related fields. Data mining, an interdisciplinary advanced technology incorporating statistics, machine learning methods, soft computing techniques, and database technologies, enables the analysis and processing of large volumes of historical data. It facilitates the extraction of hidden, unknown, and valuable information, thereby providing advanced technical support for decision-making analysis [8,9].

3.1 Research on Rainfall Prediction based on LR

Linear regression predicts continuous values, whereas LR, adapted from statistical methods for

machine learning, is used to analyze the relationship between ordered dependent variables and explanatory variables. The LR algorithm is a generalized linear regression analysis method, which uses the sigmoid function to predict the probability of events through the linear regression model. Specifically, a predicted value derived through linear regression, which is then transformed via the logistic function to convert it into a probability value, enabling predictions based on this probability. LR finds widespread applications across data mining and other fields.

$$\mathrm{z} = w^T \mathrm{x} + \mathrm{b} \tag{1}$$

$$\mathrm{a} = \sigma(z) \tag{2}$$

$$\mathrm{Y} = \mathrm{sigmoid}(wx + b). \tag{3}$$

LR's model structure can be conceptualized as a single layer of neural network, consisting of an input layer and an output layer with one sigmoid activation function, without a hidden layer. The model operation can be simplified into two steps, "linear sum of input features $[x]$ by model weight$[w]$ + sigmoid activation output probability," as shown in formulas (1)–(3).

The rainfall prediction model based on LR is constructed using multiple data dimensions derived from meteorological data. However, the original meteorological data may contain missing values and duplicate values, necessitating statistical pre-processing of the dataset. Following data pre-processing of the original data, the existing features are selected, combined with the LR model for model training, model construction is completed, and finally the model is evaluated.

3.2 Research on Rainfall Prediction based on Optimized LR

LR operates under the assumption of Bernoulli distribution and employs maximum likelihood estimation and gradient descent methods for parameter optimization to achieve predictive capabilities. The relationship between LR and multiple linear regression exhibits substantial similarities, with their primary distinction lying in the nature of their dependent variables. This fundamental commonality places both regression types within the generalized linear model family, where models share core characteristics but differ in their dependent variable distributions. Specifically, continuous dependent variables correspond to multiple linear regression, binomial distributions to LR, Poisson distributions to Poisson regression, and negative binomial distributions to negative binomial regression. Among these variants, binary LR emerges as the most widely implemented form in practical applications.

$$\mathrm{L} = -[y\log\hat{y} + (1-y)]\log(1-\hat{y}). \tag{4}$$

The learning objective of LR, derived through maximum likelihood estimation, is characterized by cross-entropy loss (also known as logarithmic loss function). This objective aims to maximize the model's predictive probability alignment with the true value distribution. Model performance improves as the predicted probability distribution approaches the true distribution more closely. It is noteworthy that LR, utilizing cross-entropy as its objective function and sigmoid activation for probability output, can only asymptotically approach probability values of 0 or 1. Consequently, the loss value can never reach exactly 0.

The optimization process employs minimized cross-entropy as its learning objective, utilizing an optimization algorithm for parameter adjustment. Given that LR under maximum likelihood estimation lacks an analytical solution, this study implements the gradient descent algorithm. Through iterative optimization, the learned parameters converge to an improved numerical solution, as illustrated in Fig. 6.

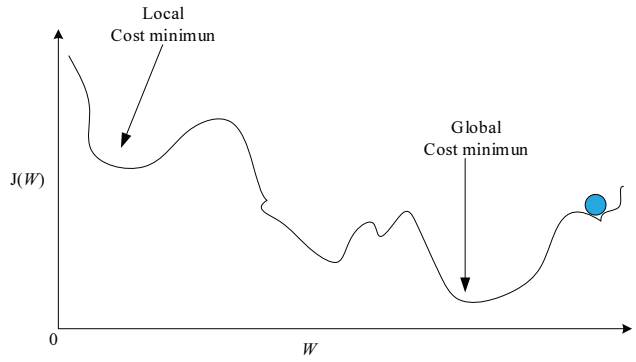


Fig. 6. Gradient descent optimization.

3.3 Research on Rainfall Prediction based on KNN

The KNN algorithm predicts new data points by calculating their distances from existing data points of various categories within the training dataset. Specifically, the algorithm makes predictions based on the KNN data points to the new observation. For distance calculation, the Euclidean distance metric is employed to evaluate the pre-processed data, as expressed in formula (5):

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{5}$$

The distance value D is obtained by calculating the difference between the new data and the historical data in the x and y dimensions, and then the distance value D is used to predict the new data.

The KNN algorithm is implemented to analyze the dataset and construct the training model. The methodology comprises the following steps: initially, data preprocessing and model construction are performed; subsequently, the trained model is applied for rainfall prediction; finally, various evaluation metrics are computed based on the prediction results to assess model performance. The process workflow is illustrated in Fig. 7.

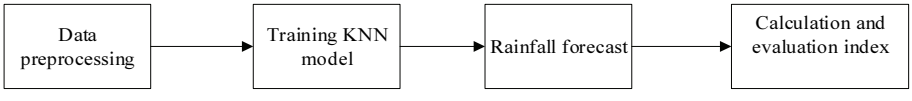


Fig. 7. Flow chart of rainfall prediction based on KNN algorithm.

3.4 Research on Rainfall Prediction based on Optimized KNN

The KNN algorithm offers several advantages: it is straightforward to implement and understand, demonstrates high accuracy, possesses well-established theoretical foundations, and is applicable to both classification and regression tasks. It effectively handles both numerical and discrete data, and proves particularly suitable for rare event classification. However, the algorithm exhibits notable limitations, including high computational and spatial complexity, intensive computational requirements, and sensitivity to class imbalance (where certain categories contain significantly more samples than others). Consequently, this algorithm is generally not recommended for large-scale datasets.

The optimal decision-making strategy for KNN involves weighted distance optimization using an inverse function formula. While the simple majority voting among k neighbors may yield suboptimal

results, as it assumes equal influence from all neighboring points, a more refined approach is necessary. This is based on the principle that the target point should share stronger similarities with nearby sample points and weaker similarities with distant ones. Therefore, the distance values require further analysis, implementing higher weights for proximate points while reducing the influence of distant points in the decision-making process.

$$f(x) = \frac{b}{(x+a)}. \tag{6}$$

As shown in formula (6), a and b can jointly control the maximum output value and change speed of $f(x)$. No matter what kind of weighting scheme is used, we should pay attention to the weighting curve not to decay rapidly, otherwise it is easy to increase the influence of noise wrongly, and cannot give enough weight to the correct sample points, so that the algorithm is too sensitive to noise, and there is a wrong conclusion that one noise point "beats" N correct samples.

3.5 Study on Rainfall Prediction based on NB

NB prediction adopts probability reasoning to provide an introduction to calculation assumptions. The prediction is based on Bayesian theorem. First, through the given training set, assuming that the eigenvalues are independent of each other, the joint probability distribution from input to output is learned. Then, based on the learned model, input X to calculate the output Y that maximizes the posterior probability, as shown in formula (7):

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}. \tag{7}$$

The NB prediction model's workflow comprises three main phases initially, meteorological data undergoes pre-processing and is partitioned into training and test sets; subsequently, the model undergoes training through data analysis and feature extraction to generate the training model; finally, the model's performance is evaluated using test data, with accuracy serving as the primary evaluation metric.

3.6 Study on Rainfall Prediction based on Optimized NB

The Bayesian algorithm encompasses three primary embedded functions: Gaussian naive Bayes (GaussianNB), polynomial distributed Bayes (MultinomialNB), and Bernoulli naive Bayes (BernoulliNB).

Gaussian distribution is also called normal distribution, a random variable X obeys the mathematical expectation μ, variance σ^2 data distribution is called normal distribution, when the mathematical expectation $\mu = 0$, variance $\sigma = 1$ is called standard normal distribution. Bernoulli distribution, also known as "zero-one distribution" and "two-point distribution," is a special case of binomial distribution. It is a special binomial distribution because it is the probability distribution of multiple Bernoulli experiments. The multinomial distribution is a generalization of the binomial distribution, where there are only two random outcome values, and the multinomial distribution where there are multiple random outcome values.

These three variants are optimized for different classification scenarios, with model selection primarily determined by data characteristics. Specifically, GaussianNB demonstrates superior performance for samples with predominantly continuous feature distributions. MultinomialNB is optimal for samples characterized by multivariate discrete features, while BernoulliNB is most suitable for samples with binary discrete features or highly sparse multivariate discrete distributions.

This study implements all three Bayesian algorithms for experimental evaluation. The experimental results presented in Section 4 demonstrate that the Gaussian Bayesian algorithm achieves optimal performance, validating that the sample features in this study predominantly follow continuous value distributions. Consequently, Gaussian Naive Bayes proves to be the most appropriate choice for this application.

3.7 Study on Rainfall Prediction based on DT

The DT algorithm leverages probability theory and employs a tree structure as its analytical framework. It utilizes decision nodes for problem-solving, represents alternative solutions through branching schemes, and models various outcomes via probability branches. Through systematic evaluation of profit and loss values across different outcome scenarios, the algorithm provides quantitative support for decision-making processes. The algorithm's construction methodology is illustrated in Fig. 8.

The DT rainfall prediction framework comprises two primary modules: the model training module and the prediction analysis module [10]. Through preprocessing and training of historical meteorological station data, a predictive model is developed. The DT model significantly enhances rainfall prediction accuracy, while the prediction analysis module applies the trained model to generate rainfall predictions for new data.

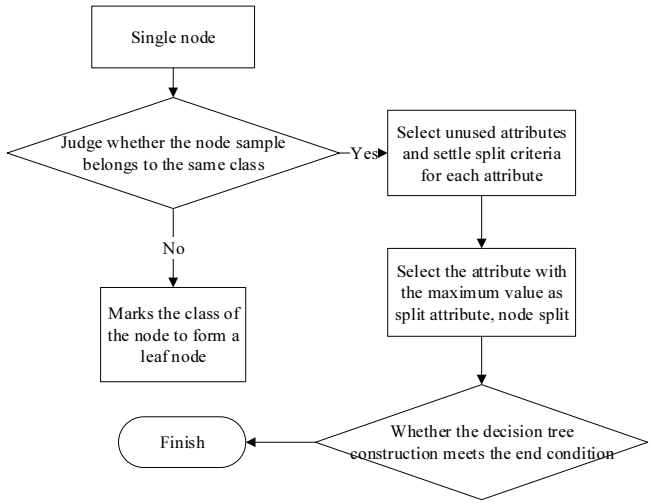


Fig. 8. DT algorithm flow chart.

3.8 Research on Rainfall Prediction based on RF

The RF algorithm is an ensemble learning method that implements a decision tree model based on the Bagging framework. The framework consists of multiple trees, each contributing to the final prediction results. The algorithm workflow is illustrated in Fig. 9, with the implementation process detailed as follows:

1) First of all, assuming that the size of the original training set is N, each tree randomly selects N samples from the training set with back as the training set of the tree, and repeats K times to form the training set of groups K;
2) Then assume that the sample dimension of each feature is m, pick a constant M, and $m < M$, and randomly select m features from M features;

3) Train each tree with m features to obtain different prediction models;
4) Finally, the optimal model is selected from (3) as the algorithm model.

The methodology initiates with the importation of meteorological data, followed by preprocessing to extract and label feature values. Subsequently, both DT and RF models are constructed and trained independently. The final phase involves the evaluation of prediction accuracy. Huang et al. [11] implemented the RF algorithm for distinguishing between sunny and rainy conditions. Their results demonstrated that the algorithm exhibited high recognition capability and achieved superior prediction accuracy.

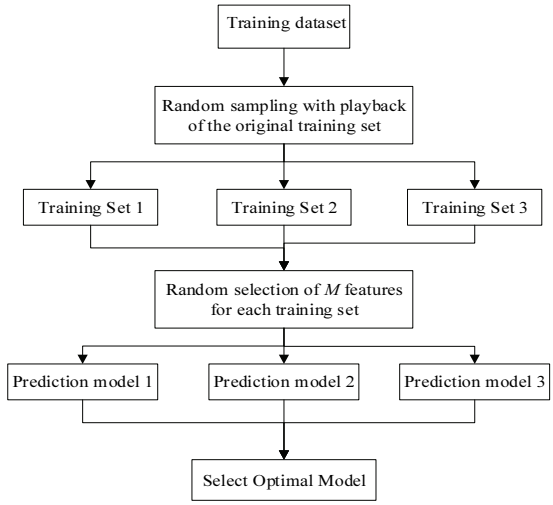


Fig. 9. Flow chart of random forest algorithm.

4. Experiment and Discussion

This study implements five machine learning algorithms: LR, KNN, NB, DT, and RF. The evaluation indexes including accuracy, error, precision, recall, and F1-score of the prediction results. The results of various prediction algorithms are obtained through in-depth research and calculation.

The results obtained after various algorithms are run are shown in Table 1.

As shown in Table 1, the DT algorithm achieves 100% accuracy on the training set but only 76.2% on the test set, indicating overfitting and demonstrating its unsuitability for this dataset. Precision, which quantifies the proportion of correctly identified positive predictions, shows that all algorithms achieve accuracy rates exceeding 76%, indicating satisfactory performance. Recall, measuring the proportion of actual positive cases correctly identified, exhibits a pattern similar to precision. Given that both precision and recall serve as indicators for positive case identification, the F1-score provides a comprehensive measure of the model's effectiveness in identifying positive cases.

The RF algorithm demonstrates superior performance in 5 out of 7 evaluation metrics. Consequently, it emerges as the optimal algorithm for this dataset. While the DT algorithm achieves 100% accuracy and zero error on the training set, these results indicate overfitting. Overfitting, characterized as model overcomplexity, typically results from the incorporation of unnecessary relationship attributes in data analysis, leading to misleading predictions.

Table 1. Rainfall prediction results of various algorithms

Evaluating index	Model				
	LR	RF	KNN	BN	DT
Train set accuracy	0.815	0.959	0.847	0.780	**1.000**
Test set accuracy	0.816	**0.837**	0.821	0.779	0.762
Train set error	0.185	0.041	0.153	0.220	**0.0**
Test set error	0.184	**0.163**	0.179	0.221	0.238
Precision	0.815	**0.835**	0.810	0.779	0.762
Recall	0.795	**0.812**	0.785	0.768	0.743
F1-score	0.815	**0.836**	0.817	0.781	0.763

The best results achieved from different methods are highlighted in bold.

Based on the findings presented in Table 1, we implement optimizations for the LR, KNN, and NB algorithms according to the methodologies proposed in Section 3. The experimental results are presented in Table 2.

LR_O and KNN_O respectively represent the optimized model. GaussianNB, MultionalNB, and BernoulliNB indicate different experimental results. As shown in Table 2, the optimized models primarily exhibit significant improvements in Precision index. Among them, LR_O algorithm improves by 1.1%, KNN_O algorithm improves by 1.2%, GaussianNB algorithm improves by 14.9% and 12.6%, respectively.

In conclusion, the optimized algorithm models demonstrate enhanced performance compared to their original counterparts, providing valuable insights and technical framework for advancing rainfall prediction methodologies.

Table 2. Comparison before and after model optimization

Evaluating index	Model						
	LR	LR_O	KNN	KNN_O	GaussianNB	MultionalNB	BernoulliNB
Precision	0.815	**0.824**	0.810	**0.820**	**0.768**	0.668	0.682

The best results achieved from different methods are highlighted in bold.

5. Conclusion

In this paper, different machine learning and optimization methods such as LR, KNN, NB, DT, and RF are used to study rainfall prediction for meteorological data. These diverse machine learning methods are systematically implemented for modeling and training, producing distinct prediction outcomes.

Rainfall prediction constitutes a crucial branch of meteorological forecasting. Addressing the issue that traditional weather forecasting fails to effectively account for the nonlinear relationship between meteorological observation data and precipitation, this paper is dedicated to applying machine learning algorithms to rainfall prediction. The research results of this paper can be summarized as follows:

1) Firstly, after reviewing extensive literature on the precipitation prediction of meteorological data, this paper chooses to analyze and predict the data with the algorithms of RF, LR, KNN, NB and DT in the machine learning algorithm;
2) Different analysis methods are used to ascertain the accuracy of rainfall prediction, and the RF algorithm has the highest accuracy.
3) The proposed traditional machine learning models are optimized accordingly, yielding improved experimental results.

Acknowledging the current limitations in knowledge and experience, future research will explore deep learning algorithms for data analysis to achieve higher prediction accuracy.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This work was supported in part by the University Level Scientific Research projects of Guangxi University of Chinese Medicine (No. 2020MS006), Guangxi University of Traditional Chinese Medicine introduced doctoral research startup Fund Project (Research on data mining algorithm of traditional Chinese medicine prescriptions #2019BS015 and Research on the Construction and Application of an AI-based Information Platform for Hepatitis B Prevention and Control in Guangxi). Thanks to the support by the National Population Health Data Center, NPHDC Population Health Data Archive.

References

[1] B. Wang, J. Lu, Z. Yan, H. Luo, T. Li, Y. Zheng, and G. Zhang, "Deep uncertainty quantification: a machine learning approach for weather forecasting," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 2087-2095. https://doi.org/10.1145/3292500.3330704

[2] I. Cvitic, D. Perakovic, M. Perisa, and B. Gupta, "Ensemble machine learning approach for classification of IoT devices in smart home," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 11, pp. 3179-3202, 2021. https://doi.org/10.1007/s13042-020-01241-0

[3] A. Al-Qerem, M. Alauthman, A. Almomani, and B. B. Gupta, "IoT transaction processing through cooperative concurrency control on fog–cloud computing environment," *Soft Computing*, vol. 24, pp. 5695-5711, 2020. https://doi.org/10.1007/s00500-019-04220-y

[4] J. Diez-Sierra and M. Del Jesus, "Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods," *Journal of Hydrology*, vol. 586, article no. 124789, 2020. https://doi.org/10.1016/j.jhydrol.2020.124789

[5] A. Alrashidi and A. M. Qamar, "Data-driven load forecasting using machine learning and meteorological data," *Computer Systems Science & Engineering*, vol. 44, no. 3, pp. 1973-1988, 2023. https://doi.org/10.32604/csse.2023.024633

[6] Chinese People's Liberation Army General Hospital, "Prostate tumor early warning dataset V1.0," 2022 [Internet]. Available: https://www.ncmi.cn/phda/dataDetails.do?id=CSTR:A0006.11.A0005.201905.000531-V1.0.

[7] M. E. A. B. Seghier, B. Keshtegar, J. A. Correia, G. Lesiuk, and A. M. De Jesus, "Reliability analysis based on hybrid algorithm of M5 model tree and Monte Carlo simulation for corroded pipelines: case of study X60 steel grade pipes," *Engineering Failure Analysis*, vol. 97, pp. 793-803, 2019. https://doi.org/10.1016/j.engfailanal.2019.01.061

[8] A. Murari, M. Gelfusa, M. Lungaroni, P. Gaudio, and E. Peluso, "A systemic approach to classification for knowledge discovery with applications to the identification of boundary equations in complex systems,"

Artificial Intelligence Review, vol. 55, no. 1, pp. 255-289, 2022. https://doi.org/10.1007/s10462-021-10032-0

[9] M. Peng, "Research and application of naive Bayesian classification algorithm in rainfall prediction," M.S. thesis, Nanjing University of Information Science and Technology, Nanjing, China, 2018.

[10] W. Yan, "The analysis and platform for meteorological data based on decision tree algorithm," M.S. thesis, Nanjing University of Information Science and Technology, Nanjing, China, 2018.

[11] X. Huang, L. Wang, S. Yang, L. Zhou, and J. Fan, "Medium-to long-term forecast of precipitation of Neijiang City based on random forest," *Yinshan Academic Journal*, vol. 31, no. 4, pp. 107-110, 2017. https://doi.org/10.13388/j.cnki.ysajs.20170628.022

Daoqing Gong **https://orcid.org/0000-0002-0977-5211**

He received the B.S. degree in computer science and technology from Guangxi Normal University, Nanning, China, in 2017. He received the master's degree of School of computer & Information Engineering, Nanning Normal University, Guangxi, China. Now, he works in School of Public Health and Management, Guangxi University of Chinese Medicine, Nanning, China. And he is pursuing a Ph.D. in Computer Science at Guangxi Normal University, Guilin, China. His research field includes evolutionary computation/deep learning/medical image processing and computational microbiomics.

Cheng Yuan **https://orcid.org/0000-0002-6365-9510**

He received the B.S. degree in information and computing science from Henan University of Engineering, Henan, China, in 2018. Now, he is a master student of School of Computer & Information Engineering, Nanning Normal University, Guangxi, China. His research field includes evolutionary computation and bioinformation processing.

Xinyan Gan **https://orcid.org/0000-0002-0341-6533**

She received the master's degree of School of computer & Information Engineering, Guangxi University, Guangxi, China. Now, she works in School of Public Health and Management, Guangxi University of Chinese Medicine, Nanning, China, as a professor. Her research field includes deep learning.

Xiang Gao **https://orcid.org/0000-0002-0546-2613**

He was born in Guilin, Guangxi, PR China, in 1989. He received the master's degree from Auburn University, USA. Now, he works in School of Public Health and Management as an associate professor, Guangxi University of Chinese Medicine, Nanning, China. His main research area is medical information.

Guizhi Sun **https://orcid.org/0000-0002-2136-9048**

She received the master's degree in signal and information processing, in 2003, and the doctor degree in 2006 form Harbin Engineering University, Harbin, China. Now she works at Guangxi University of Chinese Medicine. Her research interests are routing protocols of wireless sensor networks and wireless body area network.