Principal Component Analysis-Improved Fuzzy Genetic Algorithm

Tao Lu
Information School of Nanning University Nanning
University Nanning, China
tlu_case@126.com

Minjun Cen
Information School of Nanning University Nanning
University Nanning, China
mjdgdd@163.com

Sicong Huo*
Information School of Nanning University Nanning
University Nanning, China
huosicongedu@163.com

Ling Luo
Cultures Languages Guangxi University of Foreign
Languages Guangxi University of Foreign Languages
Nanning, China
527531662@qq.com

ABSTRACT

Since traditional intelligent algorithms composed of air quality prediction are commonly used, but such algorithms still have shortcomings for the validity of data, especially the problem of time-series prediction data. In order to investigate the problem of traditional intelligent algorithms for effective time-series data, this paper proposes a principal component analysis -improved fuzzy genetic algorithm (PCA-IFGA), in order to get more effective data in predicting air quality. the PCA-IFGA algorithm just divide into two modules. The first is PCA (Principal Component Analysis, PCA) which solves the problem of"dimensionality reduction" of air quality data by analyzing the largest individual differences revealed by taking the principal components and discovering the characteristics that affect the data in air quality prediction. The IFGA (Improved Fuzzy Genetic Algorithm) improves the traditional FGA (Fuzzy Genetic Algorithm) by increasing the variation rate of the algorithm to enhance the population diversity and facilitate the algorithm to jump out of the local optimum, while preserving the superior population diversity, increasing the crossover rate and reducing the variation rate of the algorithm to accelerate the convergence of the algorithm, thus the convergence efficiency and correct rate of the algorithm are improved. The experimental results show that PCA-IFGA is significantly better than BP algorithm, LSTM and SVM algorithms in terms of stability and full correctness.

CCS CONCEPTS

• **Mathematics of computing Information theory Coding theory**;

KEYWORDS

PCA, PCA-IFGA, Dimensionality, Reduction Introduction

ACM Reference Format:
Tao Lu, Minjun Cen, Sicong Huo, and Ling Luo. 2022. Principal Component Analysis-Improved Fuzzy Genetic Algorithm. In *2022 2nd International Conference on Control and Intelligent Robot (ICCIR 2022), June 24–26, 2022, Nanjing, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3548608.3559203

1 INTRODUCTION

In recent years, people are paying more and more attention to environmental pollution problems, such as Beijing haze, Los Angeles smog, known as the fog capital of London has also occurred smog events, people who inhale polluted air, will not only lead to bronchitis, respiratory, neurological and other system diseases or other diseases, pollutants can quickly collect in the air to a very high concentration, if humans live under such a poor air environment for a long time, the physique will also Poor air quality affects not only the human body, but also profoundly affects the environment we live in now, including deterioration of water quality, generation of acid rain, destruction of plants and their ecosystems, and reduced visibility, etc. [1]The AQI (Air Quality Index) has been established as a standard for judging the cleanliness or pollution of the air by measuring six pollutants: SO2, NO2, PM10, PM2.5, CO and O3. [2]For enterprises and society, targeted low-cost protection such as wearing masks in advance of air quality can effectively protect the health of staff, reduce the medical costs of enterprises and society, and contribute to the sustainable development of society.

Air quality prediction is a kind of time series algorithm for prediction, for example, Bhattacharya S proposed air quality prediction based on machine learning, using the advantage of fast convergence of machine learning to predict air quality, but the accuracy rate is relatively low.[3] Alsaber A R proposed random forest based air quality prediction, using the algorithm does not need to do the selection of feature values advantage, improve time efficiency, accuracy can be guaranteed to a certain extent. [4] The LSTM-based air quality prediction is proposed by Rahman A, which has less influence on noise and is relatively simple to model, but the accuracy is not high. This type of algorithm only improves the data on the model and parameters, and does not consider the validity of the air quality prediction data, so the accuracy of the predicted data is

*Corresponding author.

ICCIR 2022, June 24–26, 2022, Nanjing, China
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9717-9/22/06…$15.00
https://doi.org/10.1145/3548608.3559203

not ideal. [5]The specific impact of air quality prediction models in real life is very large, and the effect of air quality prediction can be affected by the synergy of many complex factors, such as geographic environment, human factors, climate, and human traffic. For how to obtain the data and improve the accuracy of prediction, this paper proposes principal component analysis -improved fuzzy genetic algorithm, PCA-IFGA algorithm, it firstly according to the characteristics of the original data for principal component analysis PCA, and then use IFGA algorithm for The IFGA algorithm first obtains the initial population, then it iterates on the selection of child populations to obtain new child populations, and improves the selection of child populations to increase the variability of the algorithm to solve the local optimum of the algorithm and to ensure the diversity of child populations while increasing the crossover rate of the algorithm. The crossover rate of the algorithm reduces the convergence speed of the algorithm and improves the accuracy of the algorithm.

2 PROBLEM AND ANALYSIS

Air quality prediction is a problem of time series order, and effective data in air quality prediction helps to improve the accuracy to, so this paper uses principal component analysis method to process the data, through the study of air quality data.[6] the following characteristics similarity characteristics are found through the literature: a) in the same time node similar geographical sites, the target site and the similar sites in the air quality data difference is not significant, highlighting the similarity between the geographic location and time of the data.[7] b) Through the analysis of air quality, for example, the air quality data of Beijing and the air quality data of Tianjin have certain similarity in the same time and space, highlighting the temporal and spatial similarity of the data.[8] c) Through the air quality data it is known that the air quality data in the same climate environment is similar, highlighting the the climatic similarity of air quality.[9] In summary, according to the PCA-IFGA algorithm proposed in this paper, the data of four more representative city sites, Beijing, Shanghai, Nanning, and Sanya, from September 2021-January 2022, are selected for the experiment. The analysis of the literature shows that the main pollutants affecting air quality are PM2.5, PM10, SO2, CO, NO2, and O3.

2.1 Principal Component Analysis

In this paper, the analysis of the principal component is primarily consist of two steps: (i) data cleaning of the sample data, processing the missing values in the sample data; (ii)The method as the analysis of principal component is used to reduce the dimension of meteorological data and extract the comprehensive evaluation indicators of meteorological elements. Data pre-processing: Firstly, the sample data is cleaned and the missing values in the sample data are processed. Since the amount of missing values in the sample data is a small proportion of the total sample data, the continuous missing data in the sample data are eliminated, and for a missing value with missing data, the average of the two data before and after the adjacent date is used for interpolation. The PCA-IFGA model is constructed by taking the data obtained after data cleaning as the new sample data, 80% of the sample data as the training data set, and 20% to be the validation data set.

PCA method is commonly used to deal with data with high correlation between variables. The method is a process of removing redundant information by applying the idea of dimensionality reduction and transforming multiple variables into a few principal components represented linearly by multiple variables. According to the principle of PCA, following steps are the divisions of the calculation. Standardize the data by the formulas (1) and (2), where Xij represents the standardized value of the jth variable of the ith sample; Xij represents the original data value of the jth variable of the ith sample observed in the sample data set; xj represents the initial value of the jth variable and the average value of the jth variable, σj represents the standard deviation of the jth variable.

$$\sigma_j=\frac{\sum_{j=1}^{n}(x_j-\overline{x_j})}{n-1} \tag{1}$$

$$X_{ij}=\frac{x_{ij}-\overline{x_j}}{\sigma_j} \tag{2}$$

Calculate the correlation coefficient matrix S=)sij(n*n: as in equation 3): where sij represents the correlation coefficient between the ith sample and the jth variable; xij denotes the initial value of the jth variable of the ith sample, and represents the average value of the variables of the ith and jth samples respectively.

$$s_{ij}=\frac{\left|\sum_{1}^{n}(x_{ij}-\bar{x}_j)\right|}{\sqrt{\sum_{1}^{n}(x_{ij}-\bar{x}_j)^2}} \tag{3}$$

The eigenvalues and eigenvectors of the correlation coefficient matrix are solved.

Extraction of principal components: The principal components are extracted mainly based on the value of the cumulative contribution to the sample target value Wt as in Equation 4), where yi denotes the variance of its ith principal component, zi denotes the variance of the ith sample, denoted as the total variance of the ith sample, and the larger the Wt is, the greater the contribution of its sample eigenvalues is proved, and vice versa, the smaller.

$$W_t=\frac{y_i}{\sum_{1}^{n} z_i} \tag{4}$$

2.2 Improved Fuzzy Genetic Algorithm

IFGA algorithm, like other genetic algorithms, IFGA uses individual populations and requires the acquisition of an initial population. Subsequent offspring populations are obtained by temporary populations through selection operations. In genotypes, only individual chromosomes need to be generated, and then the developmental process is controlled to carry out subsequent mutations, with continuous iterations and mutations. But the population diversity of the offspring population determines the ability of the algorithm to be locally optimal. The FGA algorithm inherits the offspring by "crossover followed by mutation", which is easy to be destroyed when selecting the best offspring and affects the timeliness of the algorithm. In this paper, we propose that IFGA only needs to randomly generate the simple individual chromosome structure of the initial population, appropriately improve the mutation rate of the algorithm and enhance the diversity of the population in the iterative process, and it is easier to jump out of the local optimization by using this feature. While obtaining the diversity of the population, IFGA algorithm can improve the crossover rate, reduce the mutation rate and improve the timeliness.

The coding structure of FGA is classified as two sections: the head can be a terminator and an operator, and the tail can only be a terminator, where L is the total length, H is the head length, T is the tail length, and M is the maximum number of operators used, as shown in the equation 5 and 6.

$$L = H + T \tag{5}$$

$$T = H\times (M - 1) +1 \tag{6}$$

The linear genotype sequence of FGA algorithm can be decoded by constructing an expression tree to obtain the corresponding expression tree, and the IFGA merit search process is based on the idea of "superiority and inferiority" of Darwin's evolutionary theory. The whole process takes chromosomes as the unit to build a population, evaluates the chromosome merit by fitness function, and generates new populations by crossover, mutation, selection and other genetic operations to iterate the merit search.

Genetic manipulation: this includes "selection", the selection of individuals with high fitness; "Crossover", through gene exchange between individuals to produce new individuals; "Mutation" refers to the generation of new individuals through gene mutations between individuals. ". genetic algorithm uses genetic operators to perform genetic operations. These operators are: selection operator, mutation operator and crossover operator.

1) Select the operator:as to the fitness of individuals, some individuals with good traits were selected from the nth generation population and inherited to the next generation (n+1) population according to certain rules. In this process of selection, the higher the fitness of the individual, the greater the chance of being selected into the next generation. For an individual i with fitness fi and population size NP, the probability formula for i to be selected is

$$P_i = \frac{f_i}{\sum_{i=1}^{N} f_i} \ (i = 1, 2, 3, 4 \ldots\ldots N) \tag{7}$$

2)Crossover operator: individuals selected in a population P(n) are randomly paired and, for each individual, some of the chromosomes (part of the coding bit string position) are exchanged between them with a specific probability (crossover probability Pc (taken as 0.25-1.0)). Crossover algorithms can better extend the search capability of genetic algorithms.

2.1) The specific steps of cross operation can be expressed as follows: 1 A pair of individuals to be mated were randomly selected from the mating pool; 2. as to the length L of the coding bit string, randomly select one or more integers K in [1, l-1] as the cross position of a pair of individuals to be mated, exchange some of their genes with each other, and form a new individual. In this paper, the chromosome evaluation function is selected and equation 2 is used as the calculation method of fitness evaluation function R2.

$$f = R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{j=1}^{m} \left((y_j - y')\right)^2}{\sum_{j=1}^{m} \left((y_j - y'')\right)^2} \tag{8}$$

where m is the number of samples, y j is the true value of the jth sample, y'j is the predicted value of the chromosome for the jth sample, and y is the mean of the sample data. the R2 value is normally within the interval [0,1], with closer to 1 meaning that the solution for that chromosome mapping is more optimal. If it is outside the interval, it means that the solution for that chromosome mapping is very different from the optimal solution and will be largely eliminated in the iterations.

3. variation operation: for each individual in the population, the gene value at one or some loci changes to other allele values with a certain probability (variation probability PM (0.01-0.1 value)). According to the individual coding method, variation can be divided into real value variation and binary variation.

3.1 Following are the variation procedure: first, judge that all individuals in the population need variation in line with the predetermined variation probability; Then, the individuals who are judged to need mutation are randomly selected for mutation.

For the course of improving the performance of IFGA Algorithm and enhance the ability of the algorithm to jump out of local optimization, this paper proposes a IFGA Algorithm that makes chromosomes adaptive. By providing adaptive crossover and mutation rates for each chromosome in each iteration, we hope to preserve the gene sequences of excellent chromosomes as much as possible and stimulate the selective potential of sub-optimal chromosomes to a greater extent. On the basis of traditional evaluation methods, this paper will increase the dominance evaluation of chromosomes in the population, and take the average value of the two as the final evaluation of chromosomes. Given a population of size N, the optimal individual fitness value of the population is Fbest, and the pros and cons of each chromosome can be evaluated by formula 9.

$$Q = \frac{F_{best}}{f' + f''} / 2 \tag{9}$$

If the value range of crossover rate and mutation rate is Min, Max, then the difference d Max Min. To ensure that good genetic sequences are not destroyed, the higher the Q value of chromosome fitness, the smaller the value of crossover rate and mutation rate. For the relatively poor chromosome, it has a higher crossover rate and mutation rate. Subsequent selection operations such as the championship can give it more opportunities to cross with the dominant chromosome in the population. In addition, the higher mutation operation probability will make it easier to produce the dominant chromosome. The specific calculation formula is as follows.

$$P_C = P_M = \text{Min} + (1 - Q) \times d \tag{10}$$

3 ANALYSIS OF EXPERIMENTAL RESULTS

In order to improve the accuracy of the algorithm, the initialized data was first subjected to a principal component analysis and the WT value that occupied the larger part of the PCA analysis was selected, as a larger WT value indicates a greater likelihood of that element influencing the prediction results. The accuracy of air quality prediction would be lower if the IFGA algorithm was used directly. Therefore, the air quality prediction is carried out after the principal component analysis. The percentages of variance and cumulative variance contribution of the six meteorological indicators data of the four cities selected in this paper after dimensionality reduction processing and principal component extraction by PCA method are shown in Table 1. IFGA related algorithm parameters are shown in Table 2

Table 1: Contribution rate of 6 meteorological index data through PCA

PCA	Initial value	variance	Cumulative contribution rate
PCA1	5.701	57.016%	58.157
PCA2	1.573	15.734%	83.264
PCA3	1.348	13.483%	86.685

Table 2: IFGA algorithm coefficient

RndSeed = random #random variable
FunctionSet = (’+’, ’-’, ’*’, ’/’, ’Q’,’S’,’O’) #Shubert
TerminalSet = (’?’, ’a’) #Terminator set
GeneNum = 4 #Number of genes
HomeGeneNum = 1 #Number of homologous genes
ConstArray = [] #Constant set
DCLen = 9 #DCDomain length
MaxArity = 2 #Maximum number of function operands
GeneHeadLen = 9 #Gene head length
HomeGeneHeadLen = 4 #Homologous gene head length
ConstArrayLen = 9 #Constant set length
m_UpperBound = 9 #Upper bound of constant set
m_LowerBound = -9 #Lower bound of constant
PopulationSize = 400 #Population size
PopulationNum = 800 #Iterations
championships = 0.01 # Tournament size
historyLen = 0 #Characteristic number

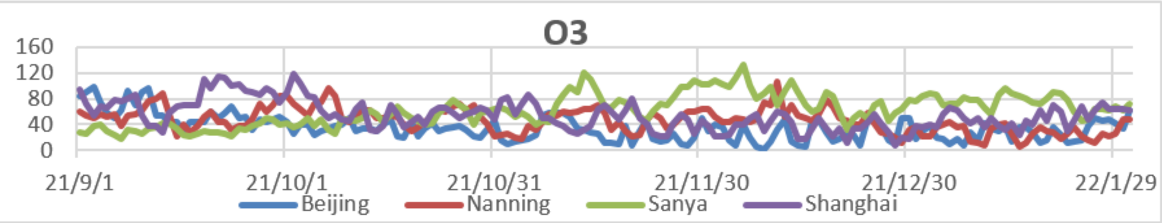

Figure 1: O3 Experimental data

In this experiment, the date set is the historical air quality data of air quality detection base stations in China. To avoid the experimental contingency caused by rainfall in the same city, the station data of four regions are used in this experiment, Sanya, Beijing, Shanghai, Nanning, respectively.[12] Due to the acquisition of the site's raw data and a large number of field names, which have site number, site latitude and longitude, the month of the data fields are not needed for this experiment, so the original data need to be pre-processed PM2.5, PM10, SO2, CO, NO2, O3 6 major pollutants concentration data, date and other fields, From the figure 1-6. Since the index of air quality in each region is different, the time period of the year when air quality is more obvious is selected as the experimental data, so the data from September 2021-January 2022 are used for this experiment.

In the above figure, the data are the air quality data of various indicators in the four regions. It can be seen from the figure that the air quality characteristics of the four regions are different from September 2020 to January 2021. In the experiment, 80% of the data samples of each data set are used as training data, 20% of the data samples are used as test data, and try to make the test data contain certain outliers. This paper mainly predicts the air quality data. In order to improve the accuracy of the experiment, considering that the air quality data is a time series, this paper will use the sliding window modeling and prediction method to model, and predict the best fitting expression through PCA-IFGA algorithm. Convert the air quality time series data processed by PCA into the matrix of formula 6, where t is the number of data in the air quality time series. Then, for any time k, the air quality value of the current time k is predicted by the air quality value of the past n times. For example, the data xn-i+1 of air quality can be predicted according

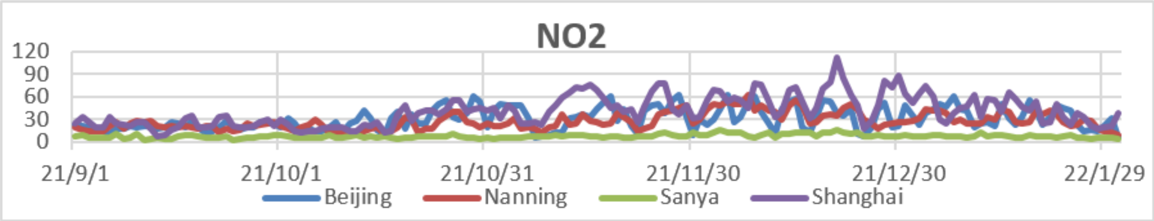

Figure 2: NO2 Experimental data

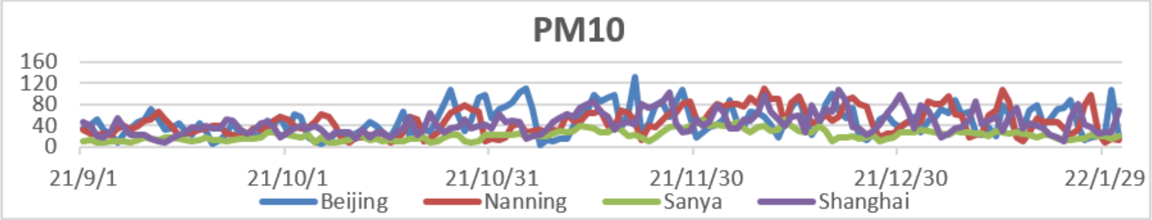

Figure 3: PM10 Experimental data

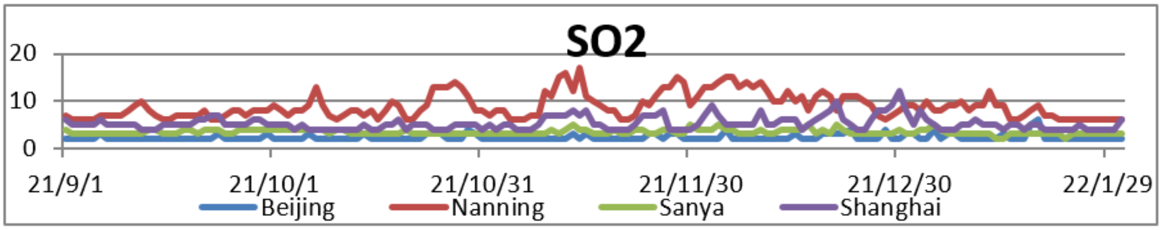

Figure 4: SO2 Experimental data

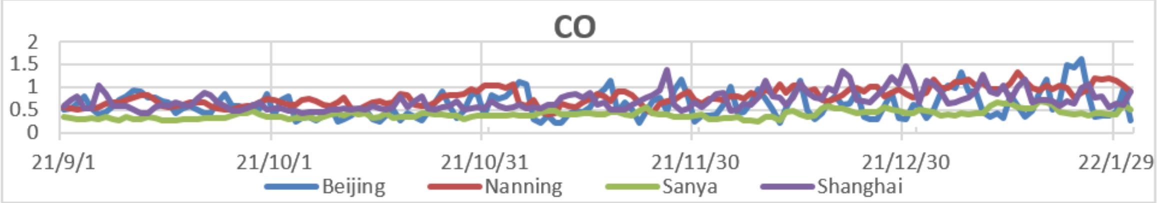

Figure 5: CO Experimental data

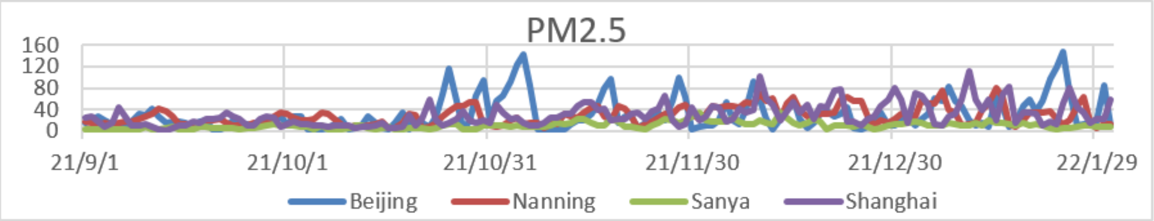

Figure 6: PM2.5 Experimental data

Table 3: PCA Coefficient

SO2	PCA1	PCA2	CO	PM2.5	PM10	PCA3	NO2	O2
0.44	0.04	0.02	0.02	0.42	0.02	0.02	0.01	0.01

to the value of Ni time in the past, such as formula 11 and 12.

$$\mathrm{Xt} = \begin{bmatrix} x_1 & \cdot s & x_{t-n+1} \\ \vdots & \ddots & \vdots \\ x_n & \cdot s & x_t \end{bmatrix} \tag{11}$$

$$x_{i-n+1}= f\,(x_1,x_2\ldots,x_{i-n}) \tag{12}$$

3.1 Experimental Environment and Evaluation Metrics

The experimental environment used in this paper is GPU:gtx1050 4GB, CPU: i5-8300h, OS: windows 11, running platform. PyCharm 2021.1.3, Python3.7. As to evaluate the air quality prediction performance of the algorithm, the traditional BP(Back Propagation ,BP) [13] and SVM (Support Vecor Machine ,SVM), LSTM(Long Short-Term Memory ,LSTM), and the prediction algorithms such as IFGA proposed in this paper were experimented on the same dataset. Among them, BP and SVM, LSTM algorithms are commonly used in air quality prediction modeling and are often used in other problems of predicting air quality. In this paper, three experimental metrics are used, which are Mean Absolute Error)MAE), Root Mean Square Error)RMSE), Accuracy)ACC).[16]

$$\mathrm{MAE} = \frac{\sum_i |\bar{x}_i - x|}{n} \tag{13}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i (\bar{x}_i - x_i)^2}{n}} \tag{14}$$

$$\mathrm{ACC} = 1 - \frac{\sum_i |\bar{x}_i - x_i|}{\sum_i x_i} \tag{15}$$

3.2 Experimental data analysis

From the table to can be seen from Table 3, it can be seen that all the nine indicators input in the PCA-IFGA model will have an impact on the city air quality indicators of Beijing, Shanghai, Nanning, Sanya Nanning, Shanghai, Beijing and Sanya, and there will be a mutual influence relationship among the indicators. Among them, SO2 and PM2.5 have a greater impact on air quality.

Summary of experimental results: From the experimental data,From the Figure 7 we can see that the predicted data of Sanya has a high fit accuracy of 96% with the real data, and there is a slight deviation in the predicted data because of the sudden change of the test set characteristics, which causes the predicted data to fluctuate. From the Figure 8-10, The prediction results of the second dataset, Shangha and Beijing, also fit correctly with the true value also have more than 98%, compared to the predicted data of Sanya and Nanning with the first two data, the data fit is not high, and the predicted rainfall data are similar to the true value error with the first two datasets. The resultant prediction accuracy is as high as 96% accuracy. The experimental data demonstrate the excellent performance of PCA-IFGA algorithm in predicting complex data such as rainfall.

From Table 4, the BP, LSTM, SVM and PCA-IFGA algorithm result metrics, it can be seen that these four algorithms predict the PCA processed air quality data of Beijing, Shanghai, Nanning and Sanya, where the mean relative error MAE and root mean square error RMSE are two common international error evaluation metrics, MAE indicates the MAE indicates the degree of dispersion of the sample data, RMSE indicates the accuracy value of the prediction, and ACC indicates the correctness of the algorithm. The MAE and RMSE of the four algorithms can be compared to verify the actual performance of their algorithmic models: the lowest value of MAE 13.68% and RMSE 17.34 for a single PCA-IFGA model reflects a measure of the degree of difference between the predicted value and the true value, the expected value of the square of the difference between the predicted value and the true value. The smaller the value of RMSE, the better the predictive ability of the PCA-IFGA prediction model, where the value of ACC is also the highest 98%, the PCA-IFGA algorithm ensures a better predictive ability while also ensuring a correct rate.

4 CONCLUSIONS

This paper is the use of PCA-IFGA algorithm for air quality prediction, the main content includes the study of the characteristics of the PCA-IFGA algorithm, the analysis of the factors affecting air quality, air quality data of the PCA and IFGA algorithm in the local optimum and convergence speed related to the performance of the improvement of the experimental results of the algorithm prediction results analysis and prediction accuracy calculation have been improved. Although the PCA-IFGA algorithm is good in predicting air quality, the algorithm as well as this experiment still has certain shortcomings and directions to be developed: (1) The basic theoretical research related to the PCA-IFGA algorithm is still lacking, such as the PCA-IFGA algorithm is not particularly ideal in parameter adjustment with randomness; (2) The PCA-IFGA algorithm is not clear enough in predictive modeling seeking optimality purpose, while innovation is needed in the cycle of the algorithm itself, and the new structure is beneficial to provide new ways of data utilization; (3) The PCA-IFGA algorithm is not ideal for crossover rate, variance rate, and population range selection, and needs further improvement.

ACKNOWLEDGMENTS

2021 Horizontal project of Nanning University. Animation rendering processing system based on image enhancement technology. Project No: 2021HX45.

Scientific Research Foundation Enhancement Project for young and middle-aged Teachers in Guangxi Universities in 2020.Agricultural products logistics loss monitoring and control system based on big data visualization technology. Project No. : 2020KY64015.

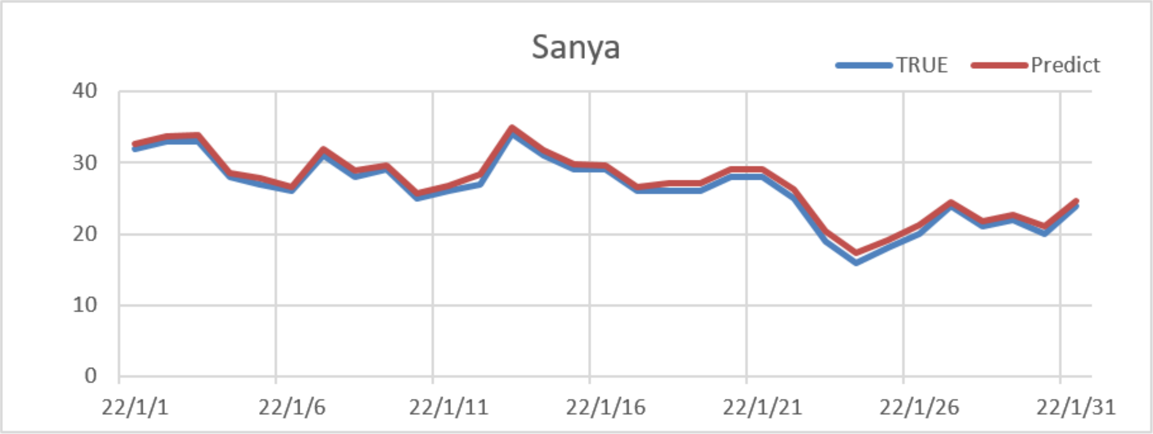

Figure 7: Comparison of Predicted and True Values in Sanya

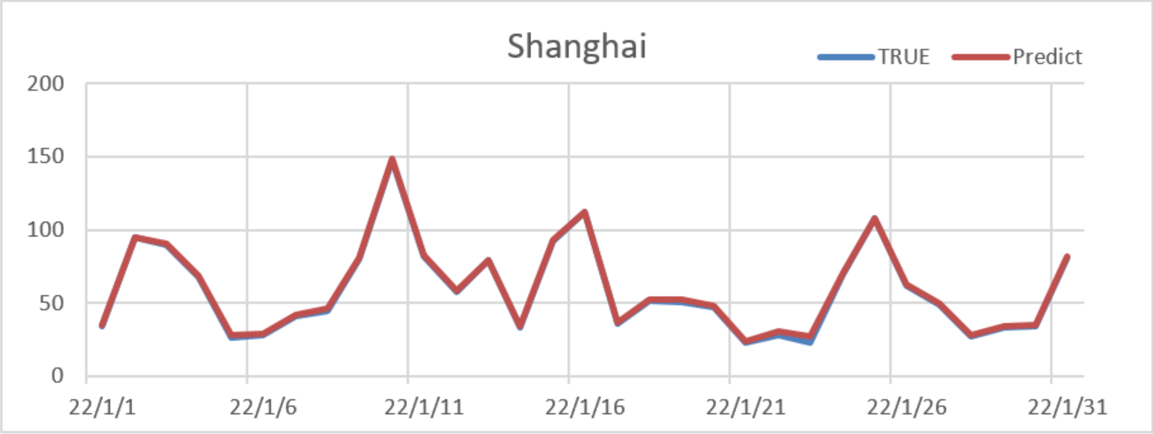

Figure 8: Comparison of Predicted and True Values in Shanghai

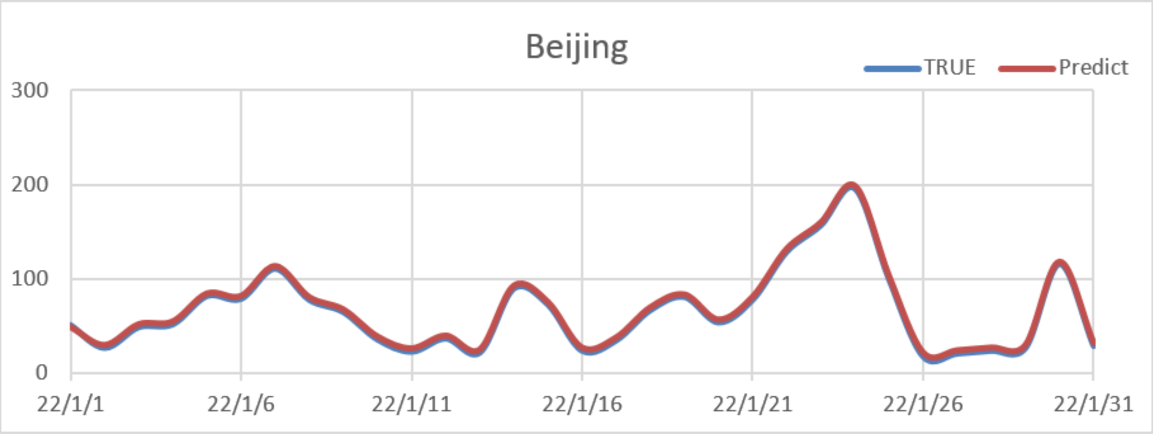

Figure 9: Comparison of Predicted and True Values in Beijing

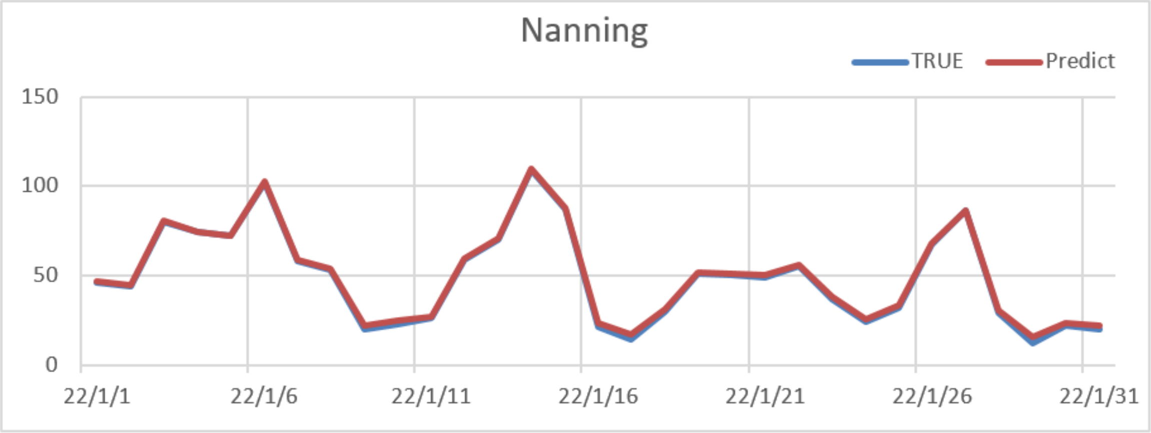

Figure 10: Comparison of Predicted and True Values in Nanning

Table 4: Algorithm metrics

Algorithm	MAE	RMSE	ACC
BP	34.45	43.02	0.802
LSTM	30.19	41.78	0.876
SVM	25.43	36.59	0.867
PCA-IFGA	13.68	17.34	0.973

The 2020 Professor Training Project of Nanning University

Application of image enhancement algorithm based on Gamma correction in temperature monitoring system. Project No. : 2020JSGC03.

REFERENCES

[1] Mackenzie A R . Introducing the Green Infrastructure for Roadside Air Quality (GI4RAQ) Platform: Estimating Site-Specific Changes in the Dispersion of Vehicular Pollution Close to Source[J]. Forests, 2021, 12.

[2] Krishnan M A , Devaraj T , Velayutham K , *et al.* Statistical evaluation of PM2.5 and dissemination of PM2.5, SO2 and NO2 during Diwali at Chennai, India[J]. Natural Hazards, 2020(9).

[3] Bhattacharya S , Shahnawaz S . Using Machine Learning to Predict Air Quality Index in New Delhi[J]. 2021.

[4] Alsaber A R , Pan J , Al-Hurban A . Handling Complex Missing Data Using Random Forest Approach for an Air Quality Monitoring Dataset: A Case Study of Kuwait Environmental Data (2012 to 2018)[J]. International Journal of Environmental Research and Public Health, 2021(3).

[5] Rahman A , Roy P , Pal U . Air Writing: Recognizing Multi-Digit Numeral String Traced in Air Using RNN-LSTM Architecture[J]. SN Computer Science, 2021, 2(1).

[6] J Kim. An Air Pollution Prediction Scheme Using Long Short Term Memory Neural Network Model[J]. E3S Web of Conferences, 2021.

[7] Kong T , D Choi, Lee G , *et al.* Air Pollution Prediction Using an Ensemble of Dynamic Transfer Models for Multivariate Time Series[J]. Sustainability, 2021, 13.

[8] Tu X Y , Zhang B , Jin Y P , *et al.* Longer Time Span Air Pollution Prediction: The Attention and Autoencoder Hybrid Learning Model[J]. Mathematical Problems in Engineering, 2021, 2021(15):1-16.

[9] Xayasouk T , Lee H M , Lee G . Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models[J]. Sustainability, 2020, 12.

[10] Silva J , Londoo L A , Varela N , *et al.* Study of the principal component analysis in air quality databases[J]. IOP Conference Series Materials Science and Engineering, 2020, 872:012030.

[11] Sitompul K L , Zarlis M , Sihombing P . Increased accuracy in the classification method of backpropagation neural network using principal component analysis[J]. IOP Conference Series: Materials Science and Engineering, 2020, 725(1):012124 (5pp).

[12] Sunori S K , Negi P B , Maurya S , *et al.* K-Means Clustering of Ambient Air Quality Data of Uttarakhand, India during Lockdown Period of Covid-19 Pandemic[C]// 2021 6th International Conference on Inventive Computation Technologies (ICICT). 2021.

[13] Xia X . Study on the application of BP neural network in air quality prediction based on adaptive chaos fruit fly optimization algorithm[J]. MATEC Web of Conferences, 2021, 336(1):07002.

[14] Hentabli M . Prediction of the concentrations of PM1, PM2.5, PM4, and PM10 by using the hybrid dragonfly-SVM algorithm[J]. Air Quality Atmosphere & Health, 2020.

[15] Chen H , Guan M , Li H . Air Quality Prediction Based on Integrated Dual LSTM Model[J]. IEEE Access, 2021, PP(99):1-1.

[16] Hancock G R , Freeman M J . Power and Sample Size for the Root Mean Square Error of Approximation Test of not Close Fit in Structural Equation Modeling[J]. Educational & Psychological Measurement, 2001, 61(5):741-758.