PM2.5 Prediction based on Multifractal Dimension and Artificial Bee Colony Algorithm

Chen Zhang, Xiaofeng Wang, Shengbing Chen, Le Zou, Chao Tang,
School of computer science and technology, Hefei University, Hefei, 230009, China
zhangchen0304@163.com

Abstract. PM2.5 pollution is becoming more and more serious in China. A novel feature selection method based on multi-fractal dimension (MFD) and artificial bee colony algorithm (ABC) is proposed to improve the prediction accuracy of PM2.5. In this method, the MFD is used as the evaluation criterion of feature subsets, and an improved ABC is taken as the search strategy. In this paper, we first use the MFD+ABC method to select the optimal feature subset, then use SVR to predict the next day’s concentration of PM2.5 in Shanghai and Guangzhou. The experimental results show that the proposed method has better performance in both the number of selected features and the prediction accuracy.

1. Introduction
PM2.5 pollution is becoming more and more serious in China, so it is necessary to monitor and effectively predict PM2.5. PM2.5 concentration is closely related to meteorological environmental data. However, meteorological environmental data contains a lot of redundant information. To improve the prediction accuracy, the optimal feature subsets should be selected before forecasting. Feature selection method can improve the accuracy and efficiency by removing redundant and irrelevant features.

Evaluation criteria and search strategy are two key issues of feature selection methods. The quality of feature subset is closely related to evaluation criteria. Mutual information [1], rough sets[2] and fractal dimension[3-4] have been used as evaluation criteria to evaluate the feature subsets. Recently, fractal feature selection methods have received more and more attention. Fractal dimension is suitable for nonlinear and high-dimension data set. Moreover, the fractal dimension can determine the number of selected features. Based on this, we use fractal dimension as evaluation criteria.

The search strategy is the other key point of feature selection. As we all known, feature selection is a discrete combinatorial optimization problem. Therefore, nature-inspired heuristic algorithms have been used as search strategies, such as ant colony optimization algorithm(ACO)[5-6] and glowworm swarm optimization algorithm(GSO)[7]. The artificial bee colony algorithm (ABC) mimics the foraging behavior of bee swarm [8]. ABC has been applied to many fields, such as the capacitated vehicle routing problem[9], the lot-streaming flow shop scheduling problem[10], wireless sensor network[11], low power FIR filter design[12]. Thus, we use an improved ABC as the search strategy.

In this paper, a novel fractal feature selection method based on the ABC and MFD is presented. The multifractal dimension is used as the assessment criterion of feature subsets, and the improved ABC is taken as the search strategy. The proposed method is utilized to forecast the PM2.5 concentration in Shanghai and Guangzhou. Compared with other feature selection methods, the proposed method is more competitive.
2. Fractal feature selection method based on MFD and the improved ABC algorithm

2.1 Multi-fractal dimension
Fractal dimension (FD) can represent the intrinsic dimension of a data set, which can determine the number of selected features[13]. Considering the complexity of the data set, multifractal dimension can better describe the characteristics of data-set compared with a single fractal dimension, which can be measured according to Equation (1)

\[
D_q = \begin{cases} 
\lim_{r \to 0} \frac{1}{q-1} \frac{\log \sum_{i} p_i^q}{\log r}, & q \neq 1 \\
\lim_{r \to 0} \frac{\sum_{i} p_i \log p_i}{\log r}, & q = 1 
\end{cases}, \quad r \in [r_1, r_2]
\]

(1)

where \( r \) is the grid size, \([r_1, r_2]\) is the scale - free interval of dataset, \( p_i \) is the probability of the data points in \( i \)th grid, \( q \) is an integer. Considering multifractal dimension’s characteristic and the distribution of datasets, \( D_{-10}, D_2 \) and \( D_{10} \) are usually used as the evaluation criteria of the degree of similarity[14].

2.2 Artificial bee colony algorithm
ABC mimics the foraging behavior of bee swarm, which consists of four components: food sources, employed bees, onlooker bees and scout bees. The pseudocode is shown as follows [15]:

1: Initialization Phase
2: Repeat
3:    Employed Bee Phase
4:    Onlooker Bee Phase
5:    Scout Bee Phase
6:    Memorize the best solution achieved so far
7: Until (Cycle = Maximum Cycle Number or a Maximum CPU time)

The details are shown as follows.

Initialization Phase
ABC generates the initial food source \( x_i = (x_{i1}, x_{i2}, \ldots, x_{iE}) \) according to Eq. (2). Each food source represents a solution.

\[
x_{ij} = x_{ij}^{\text{min}} + \text{rand}(0,1)(x_{ij}^{\text{max}} - x_{ij}^{\text{min}})
\]

(2)

where \( i = 1, \ldots, N \), \( j = 1, \ldots, E \), \( N \) is the number of food sources and \( E \) is the dimension of problem. \( x_{ij}^{\text{max}} \) and \( x_{ij}^{\text{min}} \) are the upper and lower bounds of the food source positions for the dimension \( j \) respectively. The fitness function is measured by Eq. (3)

\[
fit_i = \begin{cases} 
\frac{1}{1+f(x_i)} & \text{if } (f(x_i) \geq 0) \\
\frac{1}{1+|f(x_i)|} & \text{otherwise}
\end{cases}
\]

(3)

where \( f(x_i) \) is the fitness value of the \( i \)th food source \( x_i \), \( fit_i \) is the objective function of the food source \( x_i \).

Employed bee phase
Employed bee explores the neighborhood of the food sources, and each employed bee possesses a unique food source. The neighborhood exploration is measured by Eq. (4)

\[
v_{ij} = x_{ij} + \Phi_j (x_{ij} - x_{kj})
\]

(4)
where \( x_k \) is randomly selected from the population, which is different from \( x_i \); \( v_i \) represents the \( i \)th candidate food source; \( \Phi \) is a random number in the range \([-1, 1]\). If the candidate food source \( v_i \) is better than its parent \( x_i \), then \( v_i \) replaces \( x_i \). Otherwise, \( x_i \) remains unchanged.

**Onlooker Bee Phase**

Employed bees share the food source information with the onlooker bees. An onlooker bee chooses a food source to explored, the probability of choosing a food source is shown as Eq. (5)

\[
P_i = \frac{fit_i}{\sum_{i=1}^{N} fit_i}
\]  

(5)

The better the quality of the food source, the greater the probability of being selected. Onlooker bee searches the neighbourhood by Eq. (4). Like the employed bees, we use the greedy selection method to retain a better one from the old and new food source.

**Scout Bee Phase**

A food source will be discarded by its employed bee if it cannot be further improved over a threshold limit. Then this employed bee will become a scout bee to reproduce a new food source randomly by Eq. (2).

### 2.3 Fractal feature selection method based on MFD and the improved ABC

We propose a novel fractal feature selection method that combines the MFD and the improved ABC. In this method, MFD is used to evaluate feature subset, and the improved ABC is taken as the search strategy.

The proposed method includes the following steps:

#### 2.3.1 The fitness function

Multifractal dimension can describe the characteristics of data-set better, the fitness function is shown as Eq. (6), which is the difference of MFD between the original data-set and the feature subset

\[
f(k) = \sum_{i=1}^{N} \left( \text{frac}_q(k) - D_q \right)^2
\]  

(6)

where \( D_q \) is the \( q \)th order fractal dimension of the original data-set, which is specified with \( D_{-10}, D_2, D_{10} \), respectively[14], and \( \text{frac}_q(k) \) is the \( q \)th order fractal dimension of the selected feature subset, which is searched by the \( k \)th food source, and. The smaller the fitness value, the better the solution.

#### 2.3.2 Initialization phase

Feature selection is modeled as a discrete combinational optimization problem. Therefore, the solutions (food sources) are represented in a binary string where 0 and 1 indicate that the corresponding feature is excluded or selected, respectively.

Let \( x_i = (x_{i1}, x_{i2}, ..., x_{IE}) \) be the \( i \)th food source, which is produced randomly according to Eq. (7).

\[
x_i = (x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{IE}) \in \{0,1\}^E
\]  

(7)

where \( i = 1, ..., N, j = 1, ..., E, N \) is the number of food sources, and \( E \) is the number of feature of data-set.

For example,

\[
x = (0, 1, 1, 0, 1, 0, 0)
\]

\( x \) represents the feature subset where the second, third and fifth features are selected only.

Fractal dimension can present the intrinsic dimension, and the upper bound of fractal dimension is the number of selected features. We randomly generate an initial solution containing \( m \) ‘1’ (\( m = \lceil D_2 \rceil \)).

#### 2.3.3 Employed bee phase

Each employed bee generates a new food source \( v_i \) in the neighborhood of its parent position \( x_i \) by using the following search strategy. First, we use single-point crossover operator. The cross-over point is chosen at random. The first crossover bits of the \( i \)th candidate food source \( v_i \) are contributed by \( x_i \) and the remaining bits by \( x_k \). \( x_k \) is randomly selected, which is different from \( x_i \).
For example:

\[ x_i = (0, 1, 1, 1, 0, 1, 0, 0, 0) \]
\[ x_k = (0, 0, 1, 0, 0, 0, 1, 1, 1) \]
\[ \uparrow \text{crossover} \]
\[ v_i = (0, 1, 1, 1, 0, 0, 0, 1, 1) \]

Second, we use mutation operator. We randomly choose a number of \( E \) bits \( j \) to be flipped. Change the \( j \)th bit of \( v_i \) from 0 (1) to 1 (0).

If the candidate food source \( v_i \) is better than its parent \( x_i \), then \( v_i \) replaces \( x_i \). Otherwise, \( x_i \) remains unchanged.

2.3.4 Onlooker Bee Phase and Scout Bee Phase

In the onlooker bee phase, the neighborhood of the selected food sources is like ‘Employed bee phase’. In the scout bee phase, if a food source is discarded, the scout bee will reproduce a new food source by Eq. (7).

2.3.5 Repair the infeasible solution

In the searching process, a new solution is considered an infeasible solution if the number of “1” in the new solution is not equal to \( \lceil D_x \rceil \). Replace the infeasible solution with a new food source produced by Eq. (7).

2.3.6 Algorithm steps

The pseudo-code of the proposed is demonstrated as follows.

Initialization: Generate the \( 2N \) initial population by Eq. (7)

Divide the population into employed bee and onlooker bee by the fitness function value

While Cycle < MaxCycles do

For \( i = 1:N \) do // employed bee phase

Generate a candidate solution \( v_i \) for \( x_i \) using the method in section 2.3.3

If \( f(v_i) \leq f(x_i) \)

Replace \( x_i \) by \( v_i \)

Trial (i) = 0

Else

Trial (i) = Trial (i) + 1

End

Endfor

Calculate probability \( P_i \) by Eq. (5)

For \( i = 1:N \) do // onlooker bee phase

Select a candidate \( x_i \) according to probability \( P_i \)

Generate a candidate solution \( v_i \) for \( x_i \) using the method in section 2.3.3

If \( f(v_i) \leq f(x_i) \)

Replace \( x_i \) by \( v_i \)

Trial (i) = 0

Else

Trial (i) = Trial (i) + 1

End

Endfor

If max Trial \( \geq \) limit, // scout bee phase

Replace \( v_i \) with a randomly generated solution by Eq. (7),

Trial (max) = 0

Endif
3. PM2.5 prediction based on MFD and ABC

To improve the PM2.5 prediction accuracy, the features subset should be selected before forecasting. The meteorological environmental data have fractal characteristics [16]. Therefore, the proposed method can be used in PM2.5 prediction. First, the MFD+ABC method is used to select the feature subset (namely forecast factors). Second, SVR is used to predict PM2.5 concentration, where we use the previous day’s forecast factors to predict the next day’s PM2.5 concentration.

3.1 Simple data

Observation data sets are from January 1, 2017 to December 31, 2018 in Shanghai and Guangzhou, which are from the Shanghai and Guangzhou meteorological bureau. Meteorological data include daily average atmospheric pressure, daily maximum atmospheric pressure, daily minimum atmospheric pressure, daily average temperature, daily maximum temperature, daily minimum temperature, daily average relative humidity, daily rainfall, daily average wind speed, daily maximum wind speed and wind direction, air quality data include the daily average concentration of CO, PM10, O3, SO2, PM2.5 and NO2. These 17 influencing factors are represented by the numbers 1 to 17, respectively.

3.2 Experimental Settings

Parameters of the proposed method are given as follows: Number of food sources N=40, limit=100, Max_cycle=300.

Root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are taken as the performance measure, which are calculated according to Eqs. (7)-(10).

1. The root mean square error (RMSE)

\[
RMSE = \sqrt{\frac{1}{p} \sum_{i=1}^{p} (y_i - \hat{y}_i)^2}
\]

2. The mean absolute error (MAE)

\[
MAE = \frac{1}{p} \sum_{i=1}^{p} |y_i - \hat{y}_i|
\]

3. The mean absolute percentage error (MAPE)

\[
MAPE = \frac{1}{p} \sum_{i=1}^{p} \left| \frac{y_i - \hat{y}_i}{d_i} \right| \cdot 100\%
\]

where \( p \) is the numbers of data points, \( y_i \) is the prediction value, and \( d_i \) is the observation value, the smaller the error value, the better the prediction performance.

SVR with 3-fold cross validation (3-fold CV) is used to evaluate the experiment results. Experimental results are the average over the 3 runs of 3-fold CV.

3.3 Experimental results

Fig 1 presents PM2.5 daily average concentration from January 1, 2017 to December 31, 2018 in Shanghai and Guangzhou. The change of PM2.5 daily average concentration in Shanghai is more complex.

Table 1 and Fig. 2 show the results using MFD+ABC for Shanghai and Guangzhou. From Table 1, we find that the feature subset can provide better accuracy compared with the original data set for Shanghai and Guangzhou. The feature reduction rate of MFD+HSA is 82.35%.
Fig. 1. The daily average concentration of PM2.5 in Shanghai and Guangzhou.

Table 1 The reduction results of MFD+ABC in Shanghai and Guangzhou.

| Without feature selection | MFD+ABC | Selected features | Reduction rate (%) |
|---------------------------|---------|-------------------|-------------------|
| MAE                       | RMSE    | MAE               | RMSE              | RMSE       | Reduction    |
| 14.6771                   | 19.951  | 12.9300           | 17.6697           | 36.7675    | 82.35%       |
| 10.8492                   | 15.7065 | 8.86947           | 12.0868           | 25.4994    | 82.35%       |

Cfs, FD+GSO and FD+ABC are also used in this experiment for comparison. Cfs algorithm is implemented by WEKA. FD+GSO and FD+ABC are both fractal feature selection methods, where FD is used as the evaluation criterion, and ACO and ABC as the search strategy, respectively.

The parameters used in FD+GSO are as follows: $\rho = 0.4, \gamma = 0.6, \beta = 0.08, n_{c} = 5, l_{0} = 5, n = 50, \text{iter}_{\text{max}} = 10, r_{a}(0) = 6, r_{s} = 6, p_{1} = 0.15, p_{2} = 0.85$.

Tables 2 and 3 show the experimental results of four feature selection methods in Shanghai and Guangzhou, respectively. In terms of the number of selected features, fractal feature selection methods (FD+GSO, FD+ABC and MFD+ABC) obtained superior results. The feature reduction rate of fractal feature selection methods is 82.35% for both Shanghai and Guangzhou. However, the feature reduction rate of Cfs in Shanghai and Guangzhou are 70.59%, 76.47%, respectively. In terms of prediction errors, the MAE, RMSE and MAPE of MFD+ABC are the smallest. The errors of MFD+ABC are smaller than FD+ABC. This is because multifractal dimension can better describe the characteristic of data-set compared to single fractal dimension.

In summary, the performance of MFD+HSA are better than other methods. Furthermore, the errors of Guangzhou are smaller than that of Shanghai. This is because the concentration change of PM2.5 in Shanghai is more complicated than that of Guangzhou.

Table 2 The results of four feature selection methods in Shanghai.

| Feature selection methods | Selected feature | MAE ($\mu g/m^3$) | RMSE ($\mu g/m^3$) | MAPE (%) |
|---------------------------|------------------|------------------|-------------------|----------|
| SVR+Cfs                   | 9,12,14,15,16    | 14.6204          | 19.6126           | 48.8710  |
| SVR+FD+GSO                | 3, 4, 6          | 14.1040          | 18.8901           | 41.5037  |
| SVR+FD+ABC                | 3, 5, 9          | 13.7555          | 18.0978           | 38.5673  |
| SVR+MFD+ABC               | 3, 6, 8          | 12.9300          | 17.6697           | 36.7675  |

Table 3 The results of four feature selection methods in Guangzhou.

| Feature selection methods | Selected feature | MAE ($\mu g/m^3$) | RMSE ($\mu g/m^3$) | MAPE (%) |
|---------------------------|------------------|------------------|-------------------|----------|
| SVR+Cfs                   | 8,9,12,13        | 10.1530          | 13.8249           | 34.3699  |
| SVR+FD+GSO                | 3,19,20          | 9.86947          | 13.0868           | 31.4994  |
| SVR+FD+ABC                | 3,6,14           | 9.54805          | 12.2657           | 27.4379  |
| SVR+MFD+ABC               | 2,8,14           | 8.86947          | 12.0868           | 25.4994  |
4. Conclusion
We propose a novel fractal feature selection method that combines the MFD and the improved ABC. MFD is used as the evaluation criterion, and the improved ABC as the search strategy. The MFD+ABC method is applied to predict PM2.5 concentration. The results show that the MFD+ABC method presents better results in term of the number of selected feature and prediction accuracy.

Acknowledgments
This work is financially supported by the National Nature Science Foundation of China (Grant No. 61806068, No. 61672204), the natural science research key project of Anhui university (Grant No. KJ2018A0556, No. KJ2018A0555), the grant of Natural Science Foundation of Hefei University (Grant No. 16-17RC19, No. 0391648022), the grant of Major Science and Technology Project of Anhui Province (Grant No. 17030901026). key Technologies R&D Program of Anhui Province (Grant No. 1804a09020058).

References
[1] Liu H , Ditzler G . Speeding up joint mutual information feature selection with an optimization heuristic[C]// Computational Intelligence. IEEE, 2018.
[2] Zouache D, Abdelaziz F B. A cooperative swarm intelligence algorithm based on quantum inspired and rough sets for feature selection[J]. Computers & Industrial Engineering, 2018, 115:26-36.
[3] LópezdelPíñaa, Karmela, SolèCasals, Jordi, Egurraun H , et al. Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach[J]. Computer Speech & Language, 2015, 30(1):43-60.
[4] Zhang Chen , Ni Zhiwei , Ni Liping , et al. Feature selection method based on multi-fractal dimension and harmony search algorithm and its application[J]. International Journal of Systems Science, 2015:1-11.
[5] Harde S, Sahare V. Design and implementation of ACO feature selection algorithm for data stream mining[C]// International Conference on Automatic Control & Dynamic Optimization Techniques. 2016.
[6] Ni LiPing, Ni Zhiwei,et al. Feature selection method based on fractal dimension and ant colony optimization algorithm. Pattern Recognition and Artificial Intelligence, 2009, 22(2), 293-298(in Chinese).
[7] Ni Zhiwei, Xiao Hongwang,Wu Zhangjun, et al .Attribute selection method based on improved discrete glowworm swarm optimization and fractal dimension. Pattern Recognition and Artificial Intelligence, 2013, 26(12), 1169–1178(in Chinese).
[8] Karaboga D. An idea based on honey bee swarm for numerical optimization. Tech. Rep.-TR06. Computer Engineering Department, Engineering Faculty, Erciyes University,Turkey, (2005).
[9] Szeto, W.Y., Wu, Y.Z., Ho, S.C.: An artificial bee colony algorithm for the capacitated vehicle routing problem. Eur. J. Oper. Res. 2011, 215, 126–135.

[10] Pan, Q.K., Tasgetiren, M.F., Suganthan, P.N, et al.: A discrete artificial bee colony algorithm for the lot-streaming flow shop scheduling problem. Inf. Sci.2011, 181, 2455–2468.

[11] Yu X, Jiang J, Zhao B, et al. A self-adaptive artificial bee colony algorithm based on global best for global optimization[J]. Soft Computing, 2017(8):1-18.

[12] Dwivedi A K, Ghosh S, Londhe N D. Low power FIR filter design using modified multi-objective artificial bee colony algorithm[J]. Engineering Applications of Artificial Intelligence, 2016, 55:58-69.

[13] Encva, E., Kumaraswamy, K., & Matteucci, M.. WEKWEK: A study in fractal dimension and dimensionality reduction. Retrieved from http://home.dei.polimi.it/maneuce/Downicad/Wek kem.pdf (2002).

[14] Yan, GuangHui, Li Zhanhuai. A two phases unsupervised sequential forward fractal dimensionality reduction algorithm. Journal of Computer Research and Development, 2008, 45(11), 1955-1964 (in Chinese).

[15] Wu Liu , Sun Zhiwei , Zhang Kai , et al. Multi-population Based Search Strategy Ensemble Artificial Bee Colony Algorithm with a Novel Resource Allocation Mechanism[J]. 2017.

[16] Yuval, & Broday, D.M. Studying the time scale dependence of environmental variables predictability using fractal analysis. Environmental Science & Technology, 2010, 44, 4629-4634.