Lin *et al. EURASIP Journal on Wireless Communications and Networking* (2018) 2018:106
https://doi.org/10.1186/s13638-018-1116-3

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Air quality forecasting based on cloud model granulation

Yi Lin[1], Long Zhao[3], Haiyan Li[4] and Yu Sun[2*]

Abstract

This paper proposes a novel algorithm based on cloud model granulation (CMG) for air quality forecasting. Through data exploration of three different types of monitoring localities in Wuhan City, the determinative pollutants were reduced to NO_2, PM_{10}, O_3, and PM_{25} for modeling. After iterative granulation of original time series, the concepts of cloud model were extracted for each granule from original data space to feature space. Then, the cloud model features of future granules were predicted in the new feature space. Finally, the value in the feature space is transformed into the solution in the concept space. In addition, this paper uses the grid search to optimize the parameters in all experiments. Compare with several machine learning approaches, considering the mean squared error, the results on composition model and direct model shows that the proposed algorithm has better in predicting both individual air quality index and air quality index. At ZKX locality, the CMG algorithm can achieve high accuracy 71.43% for prediction of air quality index class. The results show that this algorithm not only can simplify the modeling process of uncertain time series in the form of knowledge abstraction, but also has good prediction performance in IAQI and AQI.

Keywords: Cloud model, Soft computing, Machine learning, Uncertainty, Air quality forecasting

1 Introduction

In health issues related to deterioration of air quality, people use air quality index (AQI) [1–3] to report on the conditions of air pollutants (APs), which were first proposed and universally acknowledged as one important determinant of adverse health effects. In the Aphekom project, the paper [4] pointed out that a reduction in APs would bring significant health and currency benefits to Europe. Recently, more and more researchers have paid attention to air quality forecasting (AQF). Previous literature has shown that the air quality forecasting is a complex problem, whose relevant data are non-linear, uncertain, and heterogeneous. Therefore, the soft computing (SC) and machine learning (ML) approaches can provide good results [1, 5–7]. When attempting to design a model combined with several APs, the problem may become more complex. Only very few attempts have been made to solve this issue [1, 8, 9]. Unfortunately, these models lost a lot of information because it converted the values of AQIs to nominal values for classification tasks.

So far, all the literature on AQF can be divided into three categories [10]: (1) simple empirical methods, (2) physically based methods, and (3) parametric or non-parametric statistical methods. In simple empirical methods, there are two ways to predict tomorrow's value: (1) using present day's value (persistence method) (2) relying critically on the dependencies between meteorological variables and predicted air pollutants. Either way, it provides low prediction accuracy. The physically based approaches model the temporal and spatial patterns of APs and meteorological variables [11], which are more accurate than the simple empirical methods. In fact, these processes are too usually too complex to be represented with physically based models. Therefore, such models can lead to biased predictions. Parametric or non-parametric statistical approaches [12, 13] can be superior to outperform physically based approaches in prediction accuracy [14], such as neural networks (NNs). However, there are some disadvantages of artificial intelligence approaches reported in the current research on APs prediction [6, 15]. Firstly, because these models are developed under meteorological conditions and chemical of some specific localities, it cannot be employed in other areas. Secondly, these models usually

* Correspondence: sunyu1225@163.com
[2]School of Computer and Electronics and Information, Guangxi University, Nanning, China
Full list of author information is available at the end of the article

simplify the meteorological process. Thirdly, the interrelationships among multiple pollutants are not modeled.

Zhang et al. [10] have shown that the NNs models give significantly higher accuracy than the linear regression approaches in the prediction of future AP concentrations, because it can model the complicated non-linear relationships between independent variables and objective functions. Recently, the support vector regression (SVR) [16] has been proven to perform better or equal than NNs in a variety of APs' predictions [17]. Compared with many other methods, such as NNs, SVR has obvious dominance. The former minimizes empirical risk and the latter minimizes structural risk. Therefore, the SVR model is relatively insensitive to the limited number of train data, and the error of test data is also limited by the SVR model. That is, the SVR model has stronger generalization performance. Lately, the fuzzy logic-based models [1, 5, 18] have been used to predict APs, which were capable of processing the inherent uncertainties in both human knowledge and data. But these approaches may suffer from computational complexity, and their prediction accuracy is lower than SVR [1].

The purpose of this study is to design one algorithm for air quality forecasting which could accommodate the uncertainties in the data at lower computational cost and good generalization performance. This algorithm not only considers the fuzziness and randomness of the problem as a whole, but also through transformation of three different state space to reduce data. First, the time series in the original data space is iteratively granulation and mapped into the new granules time series in the feature space. Then, the cloud model feature of each granule is extracted and deductive reasoning is performed. Finally, from the feature data space to the concept space, the proposed algorithm solves the original problem based on the task to be solved and the related knowledge. In addition, the paper uses the grid search to optimize all the parameters in the experiments. The results show that the proposed algorithm is capable of (1) modeling the uncertainties between data sampling points of time series; (2) maximizing generalization performance; (3) and finally costing lower computational cost.

In our study, it conducted contrast experiments at three different regions of Wuhan City. First of all, it performs data exploration to grasp the features of target data. Next, it gives the proposed algorithm's detailed description and explanation, which has also not been reported in the literature so far. Additionally, it compares this approach with several popular algorithms, non-linear autoregressive neural networks (NARNNs) [19] and SVR, on the prediction of IAQI and AQI time series. The results suggest that the CMG algorithm obtains better or nearly performance than contrast algorithms in the most experiments. At ZKX, the proposed algorithm can achieve high accuracy 71.43% on the prediction of AQI class. In the future, the research will focus on the quantitative calculation between fuzzy and random of the cloud model, and it hopes to give the value range of the problem to be solved according to the specified degree of certainty, which also can be used for other problems with uncertainty.

2 Methods

2.1 Air quality forecasting

The datasets of our studies were come from the Wuhan Municipal Environmental Protection Bureau, which contain the information about maximum daily emission variables(SO_2, NO_2, PM_{10}, CO, O_3, $PM_{2.5}$) and other knowledge (primary pollutants, Air quality index, AQI index class) from three localities in 2016~ 2017: Ganghua of Qingshan District (GH), south area of Jianghan District (JHS), and new area of Zhuankou locality (ZKX).The basic information about these studied localities is provided as follows: Ganghua locality is an urban residential zone, localized at 30°37′34.60″ north latitude, 114°22′11.13″ east longitude, altitude = 26 m; south area of Jianghan locality (JHS) is an urban commercial zone, localized at 30°18′38.86″ north latitude, 114°05′9.13″ east longitude, altitude = 16 m; new area of Zhuankou locality (ZKX) is an urban industrial zone, localized at 30°28′57.89″ north latitude, 114°09′ 24.07″ east longitude.

As presented in Table 1, there are six maximum daily emission variables, including SO_2, NO_2, PM_{10}, CO, O_3, $PM_{2.5}$, for individual localities GH, JHS, and ZK. The JHS locality has the best air quality, while the GH locality, in particular, showed high levels of PM_{10} and $PM_{2.5}$, and the ZKX locality, in particular, showed high levels of O_3. This is because the characteristics of the zones determined the air quality to a great extent [1].

After data statistical, it is not difficult to find that there is about 1.93% of data is missing. In order to improve the accuracy of air quality forecasting, there are usually three approaches to deal with missing data before modeling: (1) deleting sample instances with missing data; (2) replacing missing values; and (3) using multiple imputations. The most common way is to replace the

Table 1 Basic descriptive statistics (mean ± StdDev) of monitored pollutants

Pollutants	GH	JHS	ZKX
SO_2 [μg/m^3]	10.712±8.266	11.942 ± 7.926	13.283± 7.849
NO_2 [μg/m^3]	62.404± 23.421	65.308± 26.161	61.992 ±23.79
PM_{10} [μg/m^3]	73.617 ± 28.238	68.076 ±27.027	65.28 ± 26.536
CO [mg/m^3]	25.64±10.442	25.785 ±9.656	27.063 ± 9.703
O_3 [μg/m^3]	48.288 ± 35.028	43.277± 32.54	51.861±38.154
$PM_{2.5}$ [μg/m^3]	89.259±52.543	82.222± 43.414	79.794 ±45.31

missing value by selecting the reasonable value based on the context of the problem. In this way, it does not need to delete any sample instances. That is, this method does not lose any information but may lead to slightly biased estimates. In this paper, it uses the mean instead of missing values for numeric attribute and the most for nominal attributes.

In this study, the individual air quality index (IAQI) refers to the air quality index calculated according to the concentration of individual pollutant, whose formula is expressed as follows:

$$\mathrm{IAQI}_i = \frac{I_{\mathrm{high}} - I_{\mathrm{low}}}{C_{\mathrm{high}} - C_{\mathrm{low}}}(C_i - C_{\mathrm{low}}) + I_{\mathrm{low}} \tag{1}$$

Where C_i is the pollutant concentration of $\mathrm{AP_i}$, C_{low} denotes the nearest concentration breakpoint that is ≤ C, C_{high} denotes the nearest concentration breakpoint that is ≥ C, $\mathrm{I_{low}}$ indicates the index breakpoint corresponding to C_{low} and I_{high} indicates the index breakpoint corresponding to C_{high}. So, the IAQI formula is purely based on piecewise linear function using pollutant concentration breakpoints. In this way, the total daily health impact is expected to be the sum of the values associated with each AP [20]. In addition, the measurement intervals of these six pollutants were not identical to each other in China. For example, the concentration of $PM_{2.5}$ and PM_{10} are measured by the average of the last 24 h. SO_2, NO_2, O_3, CO are measured by the average of last hour. It is remarkable that this approach for computing AQI is different from the way used in Europe [1, 21]. Because the latter uses AQIs for the city's background and traffic conditions, it emphasizes on the role as a traffic pollutions sources.

The air quality indexes of Ministry of Environmental Protection (MEP) in China were used as target variables in this study. According to the formula published by the MEP [22], the AQI is expressed as follows:

$$\mathrm{AQI} = \max(\mathrm{IAQI}_1, \mathrm{IAQI}_2, \dots, \mathrm{IAQI}_n) \tag{2}$$

where $\mathrm{IAQI_i}$ is the individual air quality index value for the i-th air pollutant, $i = 1, 2, \dots, n$ and n denote the number of air pollutants. It is obvious that this approach uses only the maximum value of IAQI to calculate the value of AQI. In this formula, the effects of each AP are independent, without considering the interaction between of different APs. The daily maximum concentration limits of the examined air pollutants were published by China MEP as follows: SO_2: 2620 μg/m^3, NO_2: 940 μg/m^3, PM_{10}: 600 μg/m^3, CO: 60 mg/m^3, O_3: 1200 μg/m^3. While the pollutant concentration exceeds the upper limit, the maximum value of IAQI is still 500. The value beyond the maximum index is not available. In this case, it named "Beyond Index" [22]. It should be noted that the concentration limits of different air pollutants items are different from regional types and time scales. In general, the timescales are longer, the concentration limit is lower; the concentration limits of the first districts, such as nature reserves, scenic areas, and other areas requiring that need special protection, are lower than those of the second districts such as residential areas, commercial transportation, mixed residential areas, cultural areas, industrial areas, and rural areas.

As shown in Fig. 1, the boxplots depict the general span of IAQI (NO_2, PM_{10}, O_3, PM_{25}) and the AQIs in above three localities. It is not difficult to find that, the IAQI and AQI in GH locality were generally higher than other two localities. The JHS locality showed higher NO_2 and the ZKX locality showed higher O_3. In JHS locality, the AQIs were significantly lower than other two localities. The distribution range of the IAQIs and AQIs in the GH and JHS localities were similar.

In life, what people pay more attention to is the AQI class. It is more concise, intuitive, and easy to understand than AQIs. According to technical regulation on ambient air quality index [22], the AQIs were further transformed into six classes as shown in Table 2.

Figure 2 computes the cumulative days of various different air quality classes at different localities in the observation period, which shows further detailed information on the distribution of the AQI classes. The AQIs show that the air in these localities is good or slightly polluted.

Figure 3 provides the composition and percentages of determinative pollutants in the AQIs. It is found that $PM_{2.5}$ and O_3 play a determining role in the AQI computing, while NO_2 and SO_2 together account for about 20~ 30% importance. The impacts degree of determinative pollutants was calculated in the occurrence proportion of AQI.

In order to make 1 day ahead AQI predictions, our study used several machine learning algorithms. There is no one approach can do the best at any time, each approach has its own limitations and advantages. Although artificial neural networks have been widely used in various prediction domain, it has been criticized for poor generalization performance and high computational complexity [6, 15, 23, 24] Support vector regression can provide good generalization performance and lower computation cost but it may perform poorly on the noisy data. Finally, this paper proposed a regression prediction algorithm based on cloud model granulation and support vector machines, which are competent to process inherent fuzziness and randomness in both the human knowledge and data, also inherit the strong generalization performance and low running cost of SVR. The experiments with the presented methods were carried out in macOS Sierra, MATLAB R2016b Neural Network Toolbox (nntraintool), A Library for Support Vector Machines (LIBSVM-3.22) [25]. For the sake of fairness, it used grid search algorithm for parameter selection in all algorithms.

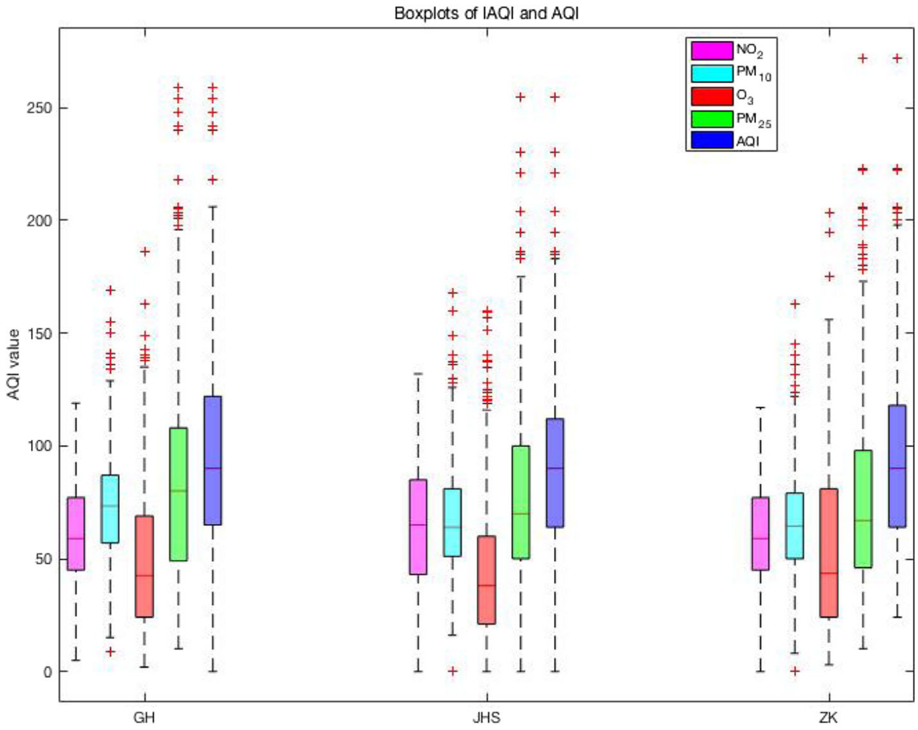

Fig. 1 Boxplots of IAQI and AQI. The boxplots depict the general span of IAQI (NO_2, PM_{10}, O_3, PM_{25}) and the AQIs in above three localities. The first group of boxplots is the distribution of these five pollutants in the GH locality. The second group of boxplots is the distribution of these five pollutants in the JHS locality. The third group of boxplots is the distribution of these five pollutants in the ZKX locality. The color of boxplot indicates the type of air pollutants

2.2 Algorithm based on cloud model granulation

Recent studies show that NNs and SVR have achieved significant improvement compared to the previous methods in the prediction of APs [12]. But the data of APs are non-linear, heterogeneous, and uncertain. The fuzzy logic systems have an advantage in dealing with the inherent uncertainty of data and human cognition, which suffer from computational cost. In this paper, it proposed an method based on concept extraction of cloud model and the conversion of state space on time series prediction of APs and AQI. This method can not only in processing uncertainty at coarser granularity, but also take into account both efficiency and generalization performance.

In 1995, the cloud model was first proposed by Deyi Li [26], who is a member of Chinese Academy of Engineering. The cloud model is a model of mutual conversion between qualitative concepts and quantitative descriptions. It has been applied in many fields, such as intelligent evaluation and fuzzy assessment. The cloud model integrates the probability theory and fuzzy set theory. By constructing a specific algorithm, the randomness, fuzziness, and relevance between concepts are unified. Cloud model does not require a priori knowledge, which can analyze the statistical rules from a large number of raw data and realize the transformation from quantitative value to the qualitative concept. It has three digital features: expectation Ex,

Table 2 AQI classes of China's Ministry of Environmental Protection

Range	AQI class	Class description
0–50	1	Excellent, no health implications.
51–100	2	Good, few hypersensitive individuals should reduce outdoor exercise.
101–150	3	Slight pollution, slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
151–200	4	Moderate pollution, slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
201–300	5	Heavy pollution, healthy people will be affected significantly. People with breathing or heart problems will experience reduced endurance in activities. These individuals and elders should remain indoors and restrict activities.
300+	6	Severe pollution, the endurance of healthy people in activities will decrease. There may be strong irritation symptoms that may cause other diseases. The old and the sick should stay indoors to avoid exercise. Healthy should avoid outdoor activities.

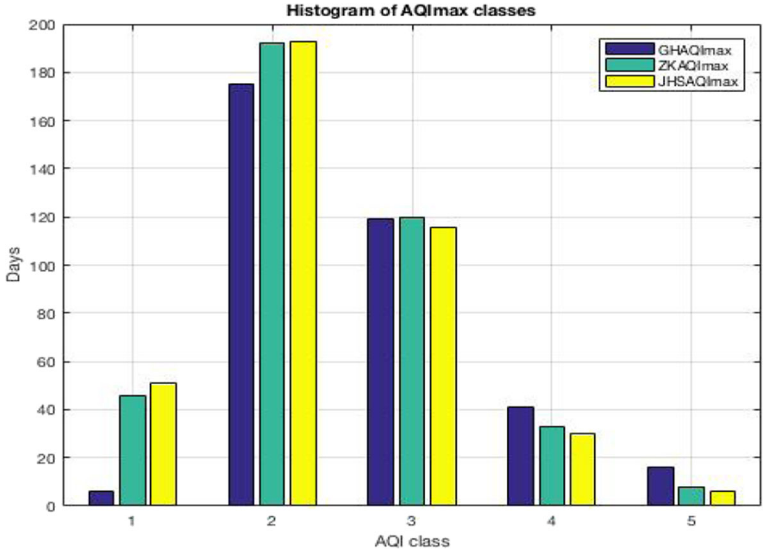

Fig. 2 Histogram of AQI classes. It computes the cumulative days of various different air quality classes at different localities in the observation period. The first group of histograms shows the days that belong to the first AQI class at three locations. The second group of histograms shows the days that belong to the second AQI class at three locations. The third group of histograms shows the days that belong to the third AQI class at three locations. The fourth group of histograms shows the days that belong to the fourth AQI class at three locations. The fifth group of histograms shows the days that belong to the fifth AQI class at three locations. The blue column is the result of the GH locality and the green column is the result of the ZKX locality. The yellow column is the result of the JHS locality

entropy En, and super entropy He, which reflects the overall features of qualitative concepts. Expectation Ex is the most representative value of the qualitative concept in the domain space. Entropy En reflects the range of number domains that can be accepted by the concept. Hyper Entropy He is the measure of entropy's uncertainty, that is entropy's entropy.

In the process of human cognitive thinking, the concepts are relative and hierarchical. Concepts or values of the same attribute usually belong to more than one upper-class concept. The proposed algorithm abstracts the target time series into mutually overlapping conceptual granules sequences then predicts them by inference judgment of qualitative concept extension. Each concept is one basic information granule, and the time range of information covered by each concept is called the window width of the granule. By backward cloud generator [27, 28], the proposed algorithm converts the distribution characteristics of data samples in each granlue into qualitative concepts represented by three digital features exception Ex, entropy En, and hyper entropy He.

After extracting concept features of the cloud model, SVR is used to predict the feature sequences Ex, En, and He respectively. The greater the probability of cloud

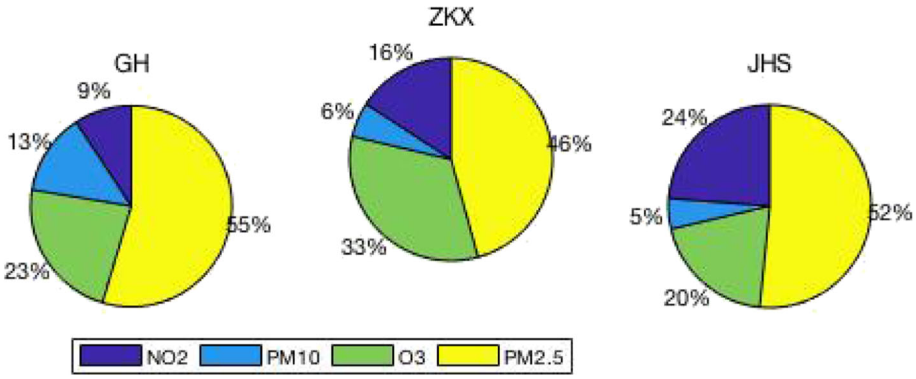

Fig. 3 Percentage of AP impact on AQIs. It provides the composition and percentages of determinative pollutants in the AQIs. The first pie is the result of the GH locality, the second pie is the result of the ZKX locality, and the third pie is the result of the JHS locality. Different colors represent different air pollutants

drop, the uncertainty of cloud drop is smaller. Therefore, the exception Ex was selected as the prediction values of the target. Specifically, the proposed algorithm based on cloud model granulation (CMG) is described in Table 3:

3 Results and discussion

3.1 Forecast results of air quality index

As mentioned above, the time series data of APs and AQIs show non-linearly characteristics over time, which are also missing, inconsistent, heterogeneous, and uncertain. It indicates that the approaches of dealing with uncertainty might perform better in air quality forecasting. In this study, it used three approaches: the support vector regression (SVR) (Vapnik, 1995), Non-linear autoregressive neural network(NAR), and the proposed SVR algorithm based on cloud model granulation, to predict the target AQI_{t+1}. In our experiments, the set of training data O_{train} (used in the procedure of learning) contained the 2016 annual data (366 data points) and the set of testing data $\mathrm{O}_{\mathrm{test}}$ from 2017.1.1~ 2017.1.7 (7 data points). In order to avoid overlearning, it used the 5-fold cross-validation to select the optimum values of the approaches' parameters.

Table 3 CMG algorithm description

```
Algorithm: CMG (TS, winSize, $n$)

Input: Time series——TS,
   Granulating window width——winSize,
   A number of days to be predicted——$n$.
Output: Qualitative predicted feature sequence of cloud model
      $\hat{E}_{\mathrm{xi}}, \hat{E}_{\mathrm{ni}}, \hat{H}_{\mathrm{ei}} (i = 1, 2, \ldots, n)$.
Algorithm steps:
A. Granulating the TS by cloud model, the digital feature sequence $E_x$, $E_n$, $H_e$ of TS is generated.
   a-1. Firstly, the original data series is converted into the granular unit data series according to the window width.
   a-2. Second, for each granular unit, the sample mean of each granular unit is calculated $\overrightarrow{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$,which is the estimated value of expectation $E_X$.
   a-3. Then, it calculates the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{X})^2$ and first order sample absolute center moments $\frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{X}|$ of each granular;
   a-4. Finally, it calculates the entropy $E_n = \sqrt{\frac{\pi}{2}} \times \frac{1}{n}\sum_{i=1}^{n} |x_i - E_X|$ and hyper entropy $He = \sqrt{S^2 - {E_n}^2}$.
B. Regression prediction of $E_x$ by SVR.
   b-1. First of all, it uses the grid search method to find the best kernel parameters for $E_X$.
   b-2. Then, it established the regression prediction model of $E_X$ by the above-selected parameter.
   b-3. Finally, it used this model to predict the expectation $Ex$.
C. Regression prediction of $E_n$ by SVR.
   c-1. First, this algorithm uses grid search method to find the best kernel parameters for $E_n$.
   c-2. Then, it established the regression prediction model of $E_n$ by the above-selected parameter.
   c-3. Finally, it used this model to predict the entropy $E_n$.
D. Regression prediction of $He$ by SVR.
   d-1. First, it uses the grid search method to find the best kernel parameters for $He$.
   d-2. Then, it established the regression prediction model of $He$ by above best parameter.
   d-3. Finally, using the model d-2 to predict $He$.
```

In this study, all the SVR approaches use the ε-insensitive loss function in the regularized risk functional that ensures optimum generalization performance. In addition, The radial basis function kernel (RBF) was used in LIBSVM toolbox [25]. In the experiments, the epsilon in loss function of epsilon-SVR is set to 0.01. The grid search method is used to find the optimum kernel parameter with penalty parameter $C = [-10, 10]$ and kernel parameter $g = [-10, 10]$.

NAR employed here is one kind of dynamic neural network, which can use to solve a non-linear time series problem with the non-linear autoregressive neural network model. The structures and parameters of the NARs were also found by a grid search method for the following values: (1) the maximum number of neurons in the hidden layer ranged was set from 10 to 100; (2) the delay days were set to $\{3, 7, 15, 90\}$ in order to the fairness of comparative experiments.

The CMG algorithm refers to the presented SVR approach based on cloud model concept extraction. Like above SVR approach, the CMG uses ε-SVR for the prediction of exception Ex, entropy En, and hyper exception He. The RBF kernel was used in sequential minimal optimization (SMOreg) for training SVR with the kernel parameter $g = [-10, 10]$ and penalty parameter $C = [-10, 10]$. According to the season effect of air quality forecasting and the physical meaning of granules, the experiments tested different values of the window width winSize=$\{3, 7, 15, 30, 90\}$;

Figure 4 presented the $\mathrm{MSEt}_{\mathrm{est}}$ of AQI prediction on both composition models and the direct models. The results indicate that the CMG algorithm trained on the direct model gets the lowest $\mathrm{MSE}_{\mathrm{test}}$ of AQI. In contrast, the highest $\mathrm{MSE}_{\mathrm{test}}$ error for the direct model was obtained using NAR. In the composition model, the lowest $\mathrm{MSE}_{\mathrm{test}}$ was also the achieved by the CMG algorithm. This observation means that, when compared with other popular artificial intelligence algorithms like SVR and NAR, the proposed CMG is able to cope with the lower amount of data in the process of learning and testing.

To further study the performance of CMG, $\mathrm{MSE}_{\mathrm{test}}$ experiments on all determinative pollutant was performed. It is obvious that the CMG algorithm performed best on the four determinative pollutants for all localities, except for the PM_{10} prediction in ZKX locality. While the Nar performed worse in most cases. The results for SVR are slightly inferior to CMG in most cases. It can be found that the high error of PM_{25} seems to generate a high $\mathrm{MSE}_{\mathrm{test}}$ in the process of AQF. It occurred despite the low errors on NO_2, PM_{10}, and O_3. The former is almost four times that of the latter in all localities. The $\mathrm{MSE}_{\mathrm{test}}$ for NO_2 was, for all methods, almost the same in the case of GH, JHS, and ZKX, respectively. For the ZKX locality, the $\mathrm{MSE}_{\mathrm{test}}$ of SVR and CMG was always lower

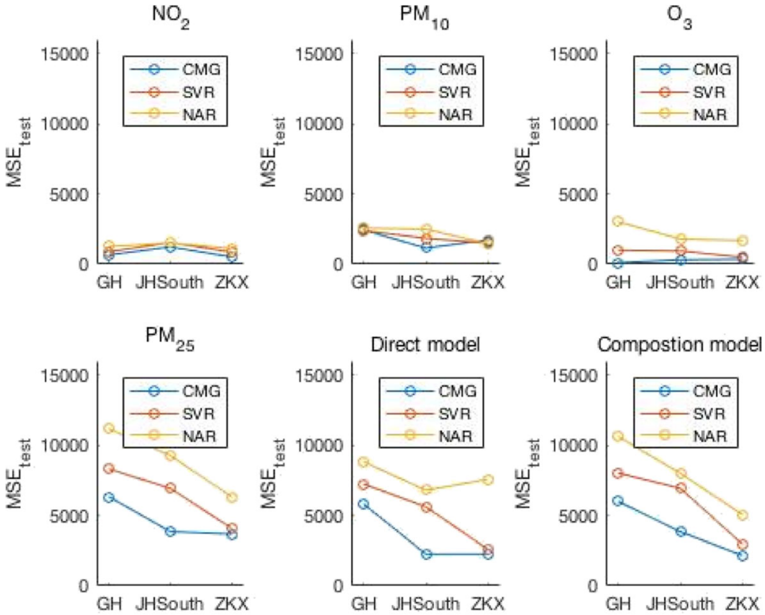

Fig. 4 MSE_{test} for AQI prediction. The first subgraph shows the MSE_{test} of three algorithms (CMG, SVR, NAR) for NO_2 prediction. The second subgraph shows the MSE_{test} of three algorithms (CMG, SVR, NAR) for PM_{10} prediction. The third subgraph shows the MSE_{test} of three algorithms (CMG, SVR, NAR) for O3 prediction. The fourth subgraph shows the MSE_{test} of three algorithms (CMG, SVR, NAR) for PM_{25} prediction. The fifth subgraph shows the MSE_{test} of three algorithms (CMG, SVR, NAR) for AQI prediction on direct model. The sixth subgraph shows the MSE_{test} of three algorithms (CMG, SVR, NAR) for AQI prediction on composition model

with the similar values on both IAQI prediction and AQI forecasting. This occurred despite the various between other methods on different localities. To sum up, these results advised that the prediction performance was related to the algorithms and the datasets. But, CMG algorithm performed best for both IAQI prediction and AQI models compared with other two algorithms.

3.2 Forecast results of air quality index class

In human life, the AQI class is the most desired, not the AQI. The prediction ability of the AQI class is one of the most interesting topics for AQI prediction model. Therefore, the above prediction values of AQI were transformed into corresponding AQI class. As shown in Fig. 5, it compared the predictive classification results of the composition models and direct models at three localities. It is not difficult to find that patterns of the test data were only presented in the 1–5 AQI class. Finally, the test accuracy of AQI class was calculated to evaluate the classification performance. Table 4 showed the percentage of correctly classified patterns of various prediction models. It shows that the direct model based on CMG for ZKX performed best up to 71.43%. The performance of CMG and SVR run better in most experiments.

As can be seen from Fig. 2, the vast majority of air quality classes can be divided into four AQI classes in three localities. In addition, the analysis of the classification results in Fig. 5 shows that the proposed CMG approaches for measuring misclassification were mainly divided into the adjacent class, i.e., from (3 → 2 or 3 → 4), from (3 → 2 or 3 → 4), from 1 → 2. Only in very few cases are divided into interval classes.

3.3 Analysis and discussion

After the data exploration, the results show that the proposed CMG approach can give better performance in both IAQI and AQI forecasting. The exploration of input data provides an understanding of the data composition of raw data. The results of the variability (mean and standard deviation) of IAQIs and AQIs indicate that the overall air quality in three localities is generally similar, but the features are locality specific. Specifically, in all localities, the percentage of the second and third classes are significantly higher than other classes, and the fifth classes have the fewest number of days. But it is worth noting that the proportion of the first class is significantly lower than the fourth class in GH, which is a contract to that in JHS and ZK. Most importantly, it is found that PM_{25} and O_3 accounted for most of the major determinative pollutants. In this study, it proposed a new algorithm based on cloud model granulation. This paper compared the performance of three approaches (nar, svr, and CMG) on both composition model and direct model in three regions.

The high importance of cloud model granulation in the prediction models gives a strategy for solving

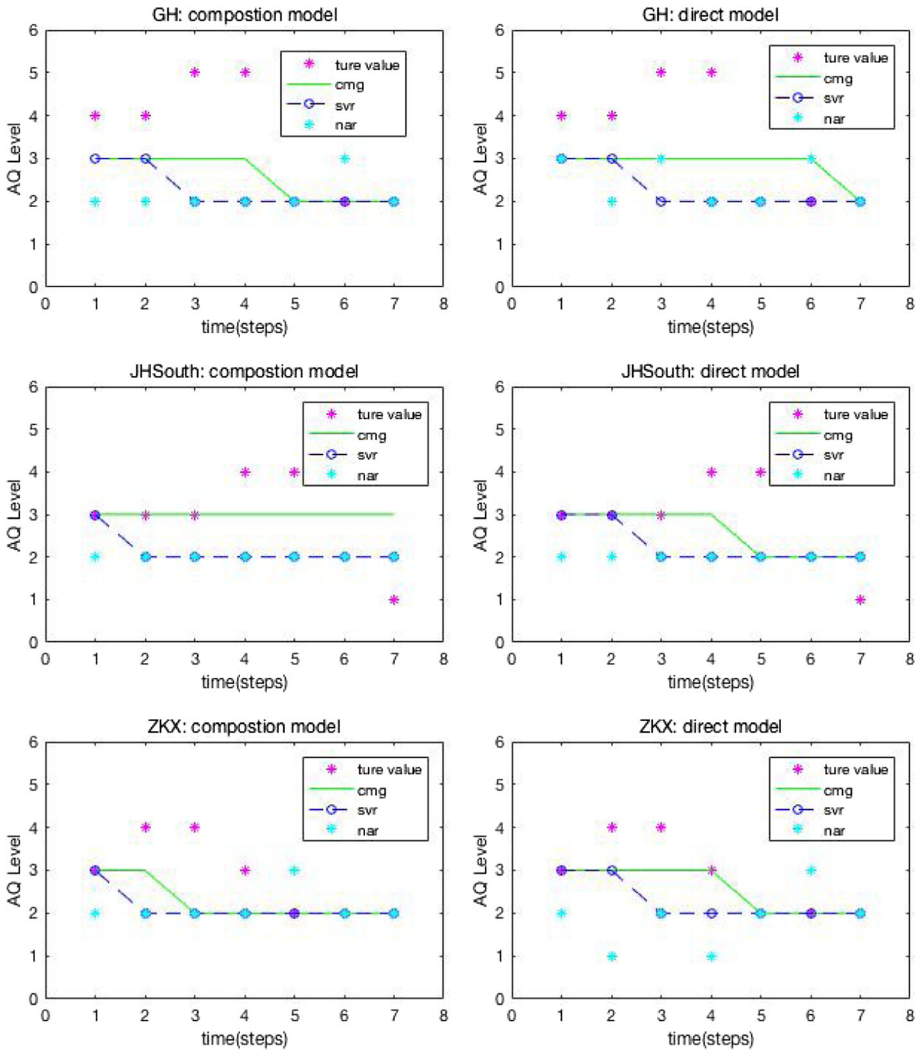

Fig. 5 Prediction of AQI class. The first subgraph shows three methods of predicting the AQL class by the composition model at the GH location. The second subgraph shows the results of three algorithms (CMG, SVR, NAR) for AQI class prediction by direct model at GH locality. The third subgraph shows the results of three algorithms (CMG, SVR, NAR) for AQI class prediction by composition model at JHS locality. The fourth subgraph shows the results of three algorithms (CMG, SVR, NAR) for AQI class prediction by direct model at JHS locality. The fifth subgraph shows the results of three algorithms (CMG, SVR, NAR) for AQI class prediction by composition model at ZKX locality. The sixth subgraph shows the results of three algorithms (CMG, SVR, NAR) for AQI class prediction by direct model at ZKX locality

problems at a higher thinking Level. It uses only three numerical features (expectation, entropy, hyper) to describe the randomness, fuzziness, and their relevance of time series data. This model enables complex data to form information granules with semantic descriptions and mine complex data in the new feature space, which is in line with human thinking and background knowledge of air quality forecasting in the real world.

In order to achieve more accurate forecasting and fair, several important artificial intelligence algorithms in AQF are compared on composition models and direct models. As expected, the designed algorithm is overall

Table 4 Prediction accuracy of AQI class

	CMG		SVR		NAR	
	Com-model	Dir-model	Com-model	Dir-model	Com-model	Dir-model
GH	42.86%	14.29%	42.86	42.86%	28.57%	28.57%
JHSouth	42.86%	42.86%	14.29%	28.57%	0	0
ZKX	57.14%	71.43%	57.14%	57.14%	28.57%	28.57%

optimal due to its ability to process uncertain information and remarkable generalization, especially in O_3 prediction. In addition, the SVR approach performs better, which corroborates the findings of a great deal of the previous work in this field of AQF [1, 17]. In the case of the AQI class' prediction, similar to AQIs forecasting, the proposed algorithm and SVR showed the promising result. The instances of misclassification are basically placed into neighboring classes.

3.4 Limitations and future work

In general, a number of important improvements of this study need to be considered. First of all, the experiments use free public data, which includes only the IAQI of APs and AQI. If more chemical and meteorological conditions information obtained, the study may get more accurate prediction results. Second, it replaced the missing data with the mean or most. It may lead to biased results. The ε-insensitive SVR with semi-supervised learning approach may use unlabeled data with missing output values. Therefore, it may be effectively used for further study. In the future, we will employ multiple output SVR for considering the error between the true value and prediction value. In addition, because classification performance was strongly influenced by class size balance, it may use the optimized classifier to solve this problem in the case of the imbalanced dataset. Last but not least, the structural model of cloud model granulation can not only enable to give a rational approach for the value prediction of IAQIs or AQIs but can also be used to estimate target's confidence interval. It hopes that the prediction range of true value can be calculated according to the given membership range, which can be used in other data with uncertain.

4 Conclusions

This paper proposed one novel algorithm based on cloud model granulation for air quality forecasting, whose data is non-linear, uncertain, and heterogeneous. After iterative granulation of original time series, the proposed algorithm extracted the conceptual features of cloud model for each new granule. Then, the cloud model feature is used as the operating space for problem solving. The CMG algorithm gives the solution to the problem by inferring the eigenvalues of future granules. The experimental results show that this algorithm can not only simplify the modeling process of uncertain time series in the form of knowledge abstraction, but also has good predictive performance in IAQI and AQI.

Compared with previous literature, this is the first attempt to predict AQIs by cloud model granulation. This method not only considers the fuzziness and randomness of the problem as a whole, but also transforms the problems in the original data space to the feature space and from feature space to concept space. Finally, the solution to the original problem is accomplished by continuous data reduction, without concern for the original data distribution characteristics. This algorithm does not only design for the time series of air quality forecasting. Its aim is to provide a new solution based on state space transformation and cloud model granulation for time series with uncertainty. There is a large research field that researchers and engineers must make more effort around data mining, machine learning, and artificial intelligence with uncertainty that these will brings light to big data. It is hoped that this will inspire readers to continue to explore cloud models to deal with the uncertainties in the real world.

Abbreviation

ML: Machine Learning; APs: Air pollutants; AQF: Air quality forecasting; AQI: Air Quality Index; GH: Ganghua of Qingshan District; JHS: South area of Jianghan District; LIBSVM: A Library for Support Vector Machines; MEP: Ministry of Environmental Protection; NARNNs: Non-linear Autoregressive Neural Networks; NNs: Neural networks; SC: Soft computing; SVR: Support vector regression; ZKX: New area of Zhuankou locality

Acknowledgements

The authors would like to thank Prof. Shuliang Wang at Beijing Institute of Technology, Beijing, China, for the help of collecting the data of this work.

Funding

This work is supported by National Natural Science Foundation of China (Grant No. 61763002), Guangxi Natural Science Foundation (Grant No. 2015GXNSFBA139249), and the Joint special of Shandong Natural Science Foundation of the China Republic under Grant No: (ZR2017LF020).

Availability of data and materials

All the experiment data is obtained from Wuhan Municipal Environmental Protection Bureau, which can be freely accessed and download on its official website.

Authors' contributions

YL and YS proposed the main idea of CMG together. YL is the main writer of this paper, wrote the code of CMG algorithm, and completed the experiments. YS analyzed the result and gave the feasible improvement plans. LZ provided the implementation solution for implementing key problems related to concept extraction by cloud model. HL simulated the transition from AQI to AQL. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Computer School, Wuhan University, Wuhan, China. [2]School of Computer and Electronics and Information, Guangxi University, Nanning, China. [3]School of Information, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. [4]School of Resource and Environmental Science, Wuhan University, Wuhan, China.

Received: 16 February 2018 Accepted: 18 April 2018
Published online: 08 May 2018

References

1. P Hajek, Predicting common air quality index—The case of czech microregions. Aerosol Air Qual. Res. **15**(2), 544–555 (2015)
2. B Bukoski, EM Taylor, Air quality forecasting. Air Qual. Manage. **87**, 563–586 (2014)
3. G Zhou, J Xu, Y Xie, L Chang, W Gao, Y Gu, J Zhou, Numerical air quality forecasting over eastern China: An operational application of WRF-Chem. Atmos. Environ. **153**, 94–108 (2017)
4. M Pascal, M Corso, O Chanel, C Declercq, C Badaloni, G Cesaroni, S Henschel, K Meister, D Haluza, P Martin-Olmedo, Assessing the public health impacts of urban air pollution in 25 European cities: Results of the Aphekom project. Sci. Total Environ. **449**, 390–400 (2013)
5. P Hájek, V Olej, Ozone prediction on the basis of neural networks, support vector regression and methods with uncertainty. Ecol. inform. **12**, 31–42 (2012)
6. N Haizum, A Rahman, MH Lee, MT Latif, Artificial neural networks and fuzzy time series forecasting: An application to air quality. Qual. Quant. **49**, 2633 (2015)
7. M El-Harbawi, Air quality modelling, simulation, and computational methods: A review. Environ. Rev. **21**, 149–179 (2013)
8. I Kyriakidis, K Karatzas, G Papadourakis, A Ware, J Kukkonen, Investigation and forecasting of the common air quality index in Thessaloniki, Greece. Artif. Intell. Appl. Innov. **382**(2), 390–400 (2012)
9. A Kumar, P Goyal, Forecasting of air quality index in Delhi using neural network based on principal component analysis. Pure Appl. Geophys. **170**, 711–722 (2013)
10. Y Zhang, M Bocquet, V Mallet, C Seigneur, A Baklanov, Real-time air quality forecasting, part I: History, techniques, and current status. Atmos. Environ. **60**, 632–655 (2012)
11. A Kumar, P Goyal, Air quality prediction of PM10 through analytical dispersion model for Delhi. Aerosol Air Qual. Res. **14**(5), 1487–1499 (2014)
12. A Donnelly, B Misstear, B Broderick, Real time air quality forecasting using integrated parametric and non-parametric regression techniques. Atmos. Environ. **103**, 53–65 (2015)
13. J Westerlund, JP Urbain, J Bonilla, Application of air quality combination forecasting to Bogota. Atmos. Environ. **89**, 22–28 (2014)
14. A-L Dutot, J Rynkiewicz, FE Steiner, J Rude, A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. Environ. Model Softw. **22**, 1261–1269 (2007)
15. W Tamas, G Notton, C Paoli, M-L Nivet, C Voyant, Hybridization of air quality forecasting models using machine learning and clustering: An original approach to detect pollutant peaks. Aerosol Air Qual. Res. **16**, 405–416 (2015)
16. M Awad, R Khanna, Support vector regression. Neural Inf. Process. Lett. Rev. **11**(10), 203–224 (2007)
17. K-P Lin, P-F Pai, S-L Yang, Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms. Appl. Math. Comput. **217**, 5318–5327 (2011)
18. D Domańska, M Wojtylak, Application of fuzzy time series models for forecasting pollution concentrations. Expert Syst. Appl. **39**, 7673–7679 (2012)
19. JT Connor, RD Martin, LE Atlas, Recurrent neural networks and robust time series prediction. IEEE Trans. Neural Netw. **5**, 240–254 (1994)
20. EK Cairncross, J John, M Zunckel, A novel air pollution index based on the relative risk of daily mortality associated with short-term exposure to common air pollutants. Atmos. Environ. **41**, 8442–8454 (2007)
21. S van den Elshout, K Léger, H Heich, CAQI common air quality index—Update with PM 2.5 and sensitivity analysis. Sci. Total Environ. **488**, 461–468 (2014)
22. Library, W. P. (2013). Ministry of Environmental Protection of the people's Republic of China.
23. G Asadollahfardi, SH Aria, M Mehdinejad, The prediction of atmospheric concentrations of toluene using artificial neural network methods in Tehran. Adv. Environ. Res. **4**, 219–231 (2015)
24. A Sadiq, A El Fazziki, J Ouarzazi, M Sadgal, Towards an agent based traffic regulation and recommendation system for the on-road air quality control. SpringerPlus **5**, 1604 (2016)
25. C-C Chang, C-J Lin, LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**, 27 (2011)
26. D Li, C Liu, W Gan, A new cognitive model: Cloud model. Int. J. Intell. Syst. **24**, 357–375 (2009)
27. C Xu, G Wang, Q Zhang, A new multi-step backward cloud transformation algorithm based on normal cloud model. Fundamenta Informaticae **133**(1), 55–85 (2014)
28. G Wang, C Xu, Q Zhang, X Wang, A multi-step backward cloud generator algorithm. Int. Conf. Rough Sets Curr. Trends Comput. **7431**, 313–322 (2012)