An Accurate and Low-cost $PM_{2.5}$ Estimation Method Based on Artificial Neural Network

Lixue Xia, Rong Luo, Bin Zhao, Yu Wang, Huazhong Yang
Dept. of E.E., Tsinghua National Laboratory for Information Science and Technology (TNList),
Tsinghua University, Beijing, China
e-mail: xialx13@mails.tsinghua.edu.cn

***Abstract*—$PM_{2.5}$ has already been a major pollutant in many cities in China. It is a kind of harmful pollutant which may cause several kinds of lung diseases. However, the existing methods to monitor $PM_{2.5}$ with high accuracy are too expensive to popularize. The high cost also limits the further researches about $PM_{2.5}$. This paper implements a method to estimate $PM_{2.5}$ with low cost and high accuracy by Artificial Neural Network (ANN) technique using other pollutants and meteorological factors that are easy to be monitored. An Entropy Maximization step is proposed to avoid the over-fitting related to the data distribution of pollutant data. Also, how to choose the input attributes is abstracted to an optimization problem. An iterative greedy algorithm is proposed to solve it, which reduces the cost and increases the estimation accuracy at the same time. The experiment shows that the linear correlation coefficient between the estimated value and real value is 0.9488. Our model can also classify $PM_{2.5}$ levels with a high accuracy. Additionally, the trade-off between accuracy and cost is investigated according to the price and error rate of each sensor.**

I. Introduction

Nowadays, the high frequency of hazy weather in many cities in China has made the particles with aerodynamic diameter less than 2.5 micrometer ($PM_{2.5}$) attract more and more attention. $PM_{2.5}$ can attach many kinds of poisonous chemicals and impact human health, which may cause many diseases such as asthma and chronic obstructive pulmonary disease (COPD) [1]. As a result, many citizens urgently want to know the $PM_{2.5}$ quality in their living and working environment.

However, the existing methods to accurately monitor $PM_{2.5}$ require the support from a high-cost and complicated system, which makes it difficult to measure $PM_{2.5}$ without a specialized monitor station [2]. It can be seen from Table I that all these highly accurate equipments need a high cost that most citizens and researchers cannot afford these equipments. As a result, monitoring $PM_{2.5}$ is far from universal, and the lack of data blocks the progress of researching and controlling of $PM_{2.5}$. Also, the high cost also leads to difficulties to analyse the $PM_{2.5}$ problem under a specific environment, such as the in-door $PM_{2.5}$ [3].

In order to reduce the cost of monitoring $PM_{2.5}$, some researchers use low-cost methods such as ANN technique to estimate $PM_{2.5}$ recently [4]–[6]. The ANN technique attempts to use data that are easy to be sensed to calculate $PM_{2.5}$. Nevertheless, $PM_{2.5}$ has complex causes and can be influenced by too many factors compared with other molecular pollutants such as O_3 [7]. The estimation accuracy is low when directly using ANN, or the cost goes high again after many kinds of expensive data are used. Given this situation, we find two

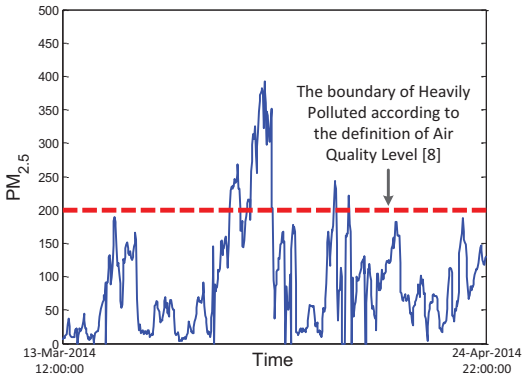

Fig. 1. IAQI of $PM_{2.5}$ over a mounth

TABLE I
Cost and Method of Equipments to Monitor $PM_{2.5}$

Method	Cost	Principle
TEOM 1405	22,000$	TEOM Gravimentric
BAM-1020	23,000$	Beta-ray
TSI DUSTTRAJ II	80,000CNY	Photometric
Dylos DC1700	425$	Particle counter

major problems that limit the estimation accuracy of ANN model and choose specific algorithms to solve them.

The first problem is the estimation error caused by the different distributions of data over different data sets. As is shown in Fig. 1, the Individual Air Quality Index (IAQI) of $PM_{2.5}$ is more likely to take a low value and only has little chance to take a high value. However, it is just the data over the boundary of Heavily Polluted range in Fig. 1 that contain important information. As a result, the important data may be ignored or only have little weights in training phase due to the small amount. This is a kind of over-fitting phenomenon which leads to a high error rate in the key range, so the trained model is inefficient when the situation of the $TestingDataset$ is different from the $TrainingDataset$, for example, the heavily polluted weeks. This paper proposes an $Entropy\ Maximization$ operation before training phase to emphasize the important data, which can avoid the over-fitting related to the data distribution and thus improve the estimation accuracy.

Second, the redundant input attributes lead to unnecessary cost and may bring noise to reduce the estimation accuracy. An attribute refers to a kind of data, for example, the $WindSpeed$. Considering that some meteorological data have an aggregation characteristic over seasons, using the irrelevant input attributes may also cause over-fitting. In fact, the problem

978-1-4799-7792-5/15/$31.00 ©2015 IEEE

can be regarded as an optimization problem whose target is estimation accuracy. This paper proposes an iterative greedy algorithm to find the better input attributes step by step.

The contributions of this work are as follows:

1) We propose an oriental low-cost method to estimate $PM_{2.5}$ based on ANN technique using the meteorological data and other pollutants. The result shows that the linear correlation coefficient R between estimated value and real value is 0.9488. The model can also be used to classify the IAQI levels of $PM_{2.5}$ with a high accuracy.
2) We propose an $Entropy\ Maximization$ step before training to normalize the distribution of data. We use information theory to analyse the problem and to provide theoretical supports for proposed algorithm. The result shows that using $Entropy\ Maximization$ can make R^2 of $TestingDataset$ increase near 0.1 while R^2 of $TrainingDataset$ decreases. So this proposed algorithm can avoid over-fitting related to the data distribution.
3) We abstract the problem of choosing input attributes as an optimization problem and propose an iterative greedy algorithm to solve it. The proposed greedy algorithm find that the $WindSpeed$ is an redundant attribute whose information is contained by others. That is, estimating $PM_{2.5}$ without $WindSpeed$ can increase R^2.
4) We create a fit curve to show the relationship between cost and estimation accuracy of the proposed model according to the price and error rate of each sensor. This curve can help user to make better decision about the input attributes in order to further reduce the cost under specific accuracy demand.

The rest of this paper is organized as follows: Section II provides related background information and related work. Section III introduces our proposed method. The results and discussions are shown in Section IV and Section V summarizes this work.

II. Preliminaries and Related Work

A. Data

According to the "Technical Regulation on Ambient Air Quality Index (AQI)" in China [8], there are six major pollutants that influence air quality, namely, CO, SO_2, NO_2, O_3, PM_{10} and $PM_{2.5}$. Considering that the other five pollutants are easy to be monitored except $PM_{2.5}$ [9]–[11], this paper uses the data of these five pollutants to estimate the IAQI of $PM_{2.5}$. IAQI is a dimensionless index ranging from 0 to 500 that describes the air quality status of individual pollutant. And the monitoring stations in Beijing provide IAQI value as a convincible high accurate data. The data of other pollutants used in this paper are the concentration of CO, and the IAQI of SO_2, NO_2, O_3 and PM_{10}. In addition, AQI is the maximum value of IAQI values of the six major pollutants above, and the air quality is divided into six levels according to AQI in order to reflect the quality more clearly, as is shown in Table II. In this paper, we use the similar level definition to classify IAQI of $PM_{2.5}$. Although the data used in this paper is from monitoring stations in Beijing and the numerical results may be regional, the proposed algorithms can be adopted by other cities.

TABLE II
Levels of Air Quality

Level	AQI range	Air Quality
1	0-50	Good
2	51-100	Moderate
3	101-150	Lightly Polluted
4	151-200	Moderately Polluted
5	201-300	Heavily Polluted
6	301-500	Severely Polluted

B. Related Work

Data mining techniques are widely used in many fields of meteorology because of their advantages such as the ability to handle big data. For example, ANN is used to estimate or predict the concentration of vehicle emission [12] and other pollutants [13], or the whole air quality [14]. However, all the researches above only use no more than two factors to estimate or to predict. Considering that $PM_{2.5}$ has a complex relationship with many factors unlike these molecular pollutants [7], the too few input attributes lead to a low estimation accuracy when dealing with $PM_{2.5}$.

Zheng uses ANN to estimate $PM_{2.5}$ [4], but the data set Zheng used only has 34 data items, so the data volume is too small to show the validity. Yao also uses ANN to estimate $PM_{2.5}$ [5], and the linear coefficient result R^2 is only 0.6556, which is not a high estimation accuracy. Also, none of these two related work consider the cost of their methods. In [6], the researchers estimate the $PM_{2.5}$ data of each place in the whole city using the $PM_{2.5}$ data from monitor stations and other data about the roads, facilities and human mobilities in the city. This work solves the $PM_{2.5}$ monitoring problem from another perspective compared with our work, so in future we can combine our method with their work to further improve the effect.

III. Proposed Method

Our estimation method is shown in Fig. 2. We first preprocess the data to reduce the error caused by sensor noise or network failure. An $Entropy\ Maximization$ operation is done before training to normalize the distribution and avoid over-fitting. Also, a greedy algorithm is proposed to iteratively find the better input attributes.

A. Preprocessing

Due to the fact that the pollutant data have a low sample rate with high fluctuation, it is difficult to identify whether a data item contains wrong point or not. Given this situation, this paper uses specific method to eliminate wrong points. A element is considered as a wrong point when its value fluctuate fiercely while the others are stable. Nevertheless, the data cleaning approach above produces many blank elements in the data set. As for ANN, the number of input attributes can't be changed in a certain system. Therefore, we use liner interpolation method to fill up the blank elements.

The reconstruction error of liner interpolation is small enough compared with the large amount of data. Also, liner interpolation is easy to be implemented and can be computed at high speed, which meets the need of our system.

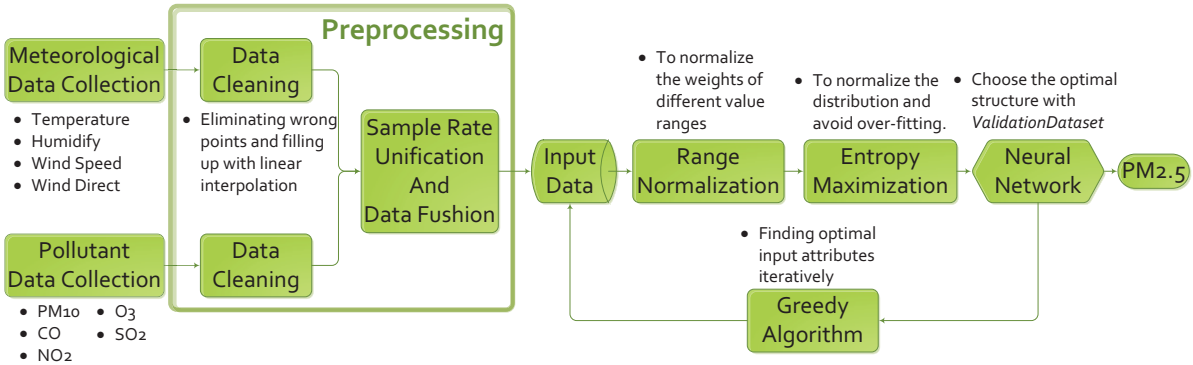

Fig. 2. Our accurate and low-cost $PM_{2.5}$ estimation method framework

B. Entropy Maximization

It is always just the rare data values that contain some important information, such as the heavy pollution shown in Fig. 1. So the important data may be ignored or only have little weights in training phase. As a result, the trained model may have a larger error rate in the key part, which is caused by the over-fitting related to the data distribution. As for the whole data set, it is impossible to guarantee that the $TrainingDataset$ has a same or approximate distribution with any practical data set collected in future. So this over-fitting phenomenon limits the estimation accuracy of the trained model.

For example, as for our $TrainingDataset$ that contains 6801 data items, there are 256 $PM_{2.5}$ data whose value is in the severely polluted range, namely, the range of values larger than 300. So in the training phase, the severely polluted data only has a little chance, nearly 3%, to train the model. Therefore, the trained model may have a large error rate in severely polluted range. And the estimation accuracy can be very low when we use the trained model to estimate $PM_{2.5}$ in severly polluted weeks whose $PM_{2.5}$ values are all larger than 300.

We consider that the inefficient result of the example above is similar to an over-fitting phenomenon. This over-fitting phenomenon is caused by the obviously different distributions between $TrainingDataset$ and $TestingDataset$. Therefore, we want to normalize the distributions before training phase to avoid over-fitting.

We find the relationship between distribution and information can be explained by the concept of $Information\ Entropy$ in information theory, and some theorems can provide theoretical support for the normalization of distribution. Given a discrete random distribution $P = \{p_1, p_2, ...p_n\}$, the $entropy$ of the distribution P is defined by [15]:

$$H(P) = \sum_{i=1}^{n} p_i \log(\frac{1}{p_i}) \tag{1}$$

where $\log(\frac{1}{p_i})$ is the $individual\ information$ of a random event that have a probability p_i to occur. So the event with lower probability to occur contains larger information, which is similar to the information of data. The $entropy$ of a random variable with distribution P can reflect the information of the variable itself, which is the weighted average of each value's information. The $entropy$ of the random variable reaches a maximum value when P equals to a uniform distribution, that is [15]:

$$0 \leq H(P) \leq \sum_{i=1}^{n} \frac{1}{n} \log(n) = \log(n) \tag{2}$$

Considering the finite precision, the value of an input attribute can also be regarded as a discrete random variable when the data amount is large enough. So we can maximize the entropy if we normalize the distribution of each attribute into a uniform distribution or an approximately uniform distribution according to Eq.(2), this normalization of distribution is defined as $Entropy\ Maximization$ step. After $Entropy\ Maximization$, the information of each attribute can be fully utilized to improve the effect of the model. And because the distribution has been normalized, we can get a more general model over different data distributions, which can avoid over-fitting and further increase the estimation accuracy.

In order to implement the $Entropy\ Maximization$, this paper uses the existing conclusion from probability theory to normalize the distribution after the range normalization step shown in Fig. 2. If we already have a prior distribution knowledge of a random variable x whose distribution is $p(x)$ with range of interval $[0, 1]$, we can normalize x into a random variation y with uniform distribution using the function:

$$y = g(x) = \int_0^x p(x)dx \tag{3}$$

Here we use the distribution of train data as the prior distribution, so the distribution can be normalized by Eq.(3). This function is a monotonic non-decreasing function and is not sensitive to noise. However, the $Entropy\ Maximization$ changes the definition of distance between different samples, but ANN technique is not strictly based on a distance metric, which makes the $Entropy\ Maximization$ method suitable for our method. The comparison between our method and the experiment without $Entropy\ Maximization$ is shown in Section IV.D

C. Iterative Greedy Algorithm to Choose Input Attributes

Since we don't know the inner relationship between input attributes, the component of input attributes is difficult to decide. If we use as more attributes as possible, not only the cost can be relatively high, but also the estimation accuracy

Algorithm 1: Greedy Algorithm to Choose Input Attributes

Input: $OriginalInputAttributes$, $TrainingDataset$, $ValidationDataset$,$NetworkParameters$
Output: $OptimalInputAttributes$

$flag \leftarrow true$
$CurrentAttributes \leftarrow OriginalInputAttributes$
Initial the Network with $NetworkParameters$ and number of $CurrentAttributes$;
Train the Network with $TrainingDataset$ and $CurrentAttributes$;
Evaluate the result with $ValidationDataset$ to get $CurrentAccuracy$;
while $flag == true$ **do**
  $flag \leftarrow false$
  for $input \leftarrow CurrentAttributes$ **do**
    $TempAttributes \leftarrow CurrentAttributes - input$
    Initial the Network with the number of $TempAttributes$;
    Train the Network with $TempAttributes$;
    Evaluate the result to get $accuracy$;
    if $accuracy > CurrentAccuracy$ **then**
      save $input \rightarrow InputBuff$
      save $accuracy \rightarrow AccuracyBuff$
      $flag \leftarrow true$
    end
  end
  $input \leftarrow$ arcmax($accuracy$)
  $CurrentAttributes \leftarrow CurrentAttributes - input$
end
$OptimalInputAttributes \leftarrow CurrentAttributes$

may decrease. The redundant attribute may introduce noise into the model and cause over-fitting. There are two reasons. First, the combination of some input attributes may already contains the information of a redundant attribute. Second, some attributes may have little relationship with $PM_{2.5}$, so introducing the irrelevant attributes may establish an incorrect causality.

In fact, the choice of suitable input attributes can be regarded as an optimization problem. The adjustable parameter is the kinds of input attributes, and the optimization target is estimation accuracy of final model. We propose an iterative greedy algorithm to solve the optimization problem and find the most suitable input attributes. The steps of the proposed algorithm is shown in Algorithm 1.

In the proposed greedy algorithm, we delete each input attribute in turn and test the estimation accuracy of the model trained by other input attributes. We can find some attributes from the testing result that when we delete them, the accuracy increases on the contrary. We greedily eliminate the attribute that limits the accuracy most in each iteration and find the better input attributes step by step. Additionally, we introduce a $ValidationDataset$ independent of $TrainingDataset$ to examine each subset's accuracy. The $ValidationDataset$ is also independent of the final $TestingDataset$, so the operation of choosing input attributes doesn't lead to the over-fitting result.

However, there is also a trade-off between cost and estimation accuracy. After deleting redundant and irrelevant attributes, we can choose fewer input attributes for lower cost but the accuracy decreases at the same time. In order to determine which choice is better considering to both cost and accuracy, we calculate the cost of different choices using the price of sensors and then create a fit curve for decision. More details are discussed in Section IV.D and Section IV.E.

D. Structure and Parameters of BP-ANN

We use Back Propagation Artificial Neural Network (BP-ANN) to implement the estimation model, where the only one output neuron calculates the $PM_{2.5}$ result of an input data item. If we want to regard our goal as a classification of $PM_{2.5}$'s IAQI levels instead of an estimation of $PM_{2.5}$'s IAQI itself, we only need to change the output layer of BP-ANN structure to meet the classification demand. That is, considering that there are six levels, the output layer should have six nodes, and each node describe whether the $PM_{2.5}$'s IAQI of current data item is in the corresponding level.

There are also many parameters that need to be determined before training phase such as the number of neurons in hidden layer. We use the $ValidationDataset$ mentioned before to examine result of the parameter's each value and choose the value that can lead to the highest estimation accuracy.

IV. Experiment Results and Discussions

In this section, we firstly introduce the evaluation criteria and data set in our work. Then the final estimation and classification result is shown with the comparison of related work. In order to illustrate the effect of our proposed algorithms, we also give the result of different choices about structure and input attributes. Finally, a fit curve reflecting the trade-off between the cost and accuracy is provided.

A. Evaluation Criteria

Here we use three factors to show the relevance and accuracy of different meaning, which can reflect the effect of our method from different angles:

1) The linear correlation coefficient R between estimated values and real values. Since the estimated values should equal the real values ideally, namely, $R = 1$, R can reflect the correlation between estimated values and real values from a global scale. Also, R is an evaluation criterion generally used in meteorology, which makes it easy to compare our work with others.
2) The number and proportion of estimated IAQI values classified in the same level with real value. Although the level is defined on AQI, we can extend this definition to IAQI to provide a visualized classification accuracy of discrete meaning.
3) The number and proportion of estimated IAQI values whose difference from the corresponding real value is not larger than 50. Considering that the size of an IAQI level′s range can be 50, 100 or 200, this evaluation criterion can be regarded as a classification accuracy of continuous meaning, which means the estimated value does not deviate its real value more than an index size.

B. Data Set and Input Attributes

In this work, we use the data from 13 monitoring stations in Beijing during 684 hours (about 4 weeks) to train the network, and the $ValidationDataset$ includes data of 329 hours (about 2 weeks). We use the data from 08 am 4/3/2014

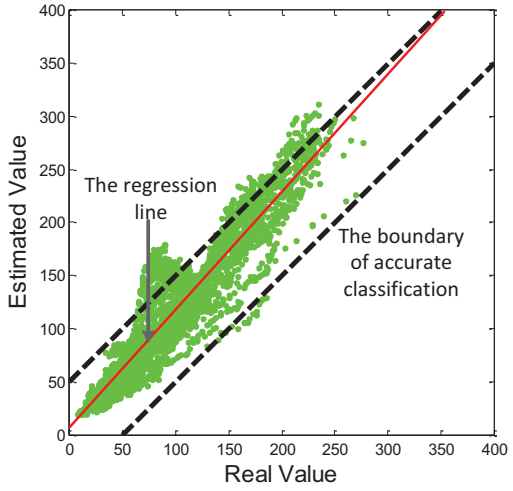

Fig. 3. Estimation result and accurate range

TABLE III
COMPARISON WITH RELATED WORK

Work	Pollutant	Data Amount	R^2 with ANN
[16]	O_3	A Week	0.845
[5]	$PM_{2.5}$	5 days($<$ 200)	0.6556
Ours	$PM_{2.5}$	2 weeks(3686)	0.9002

TABLE IV
CLASSIFICATION RESULT

Number		Classification Level						Recall (%)
		1	2	3	4	5	6	
Real Level	1	977	78	0	0	0	0	93
	2	160	719	176	135	4	0	60
	3	0	124	403	156	0	0	59
	4	0	1	41	352	144	0	65
	5	0	0	0	6	180	2	96
	6	0	0	0	0	3	25	89
Precision(%)		86	78	65	54	54	93	73

to 10 am 4/17/2014 during 323 hours to test our method. The $TestingDataset$ has total 3686 valid data items and the $TestingDataset$ is independent of $TrainingDataset$ and $ValidationDataset$. The initial data set has 9 attributes including $Temperature$, $Humidity$, $WindSpeed$, $WindDirection$, CO, SO_2, NO_2, O_3 and PM_{10}, and the $WindSpeed$ is recognized as a redundant factor which is not used in the final model.

C. Estimation and Classification Result

The final result is shown in Fig. 3. Each point refers to an estimation result whose x-coordinate is the real value and y-coordinate is the estimated value calculated by ANN. The linear correlation coefficient R of $TestingDataset$ is 0.9488. It can be seen from Fig. 3 that the points have an obvious linear convergent feature, which means our method can reach a great estimation result for data set with a large volume.

The comparison between our work and others is shown in Table III. [16] uses ANN and principal component regression to predict O_3 , and [5] uses ANN to estimate $PM_{2.5}$. It can be seen from Table III that our result achieves a better effect and can handle a larger data set.

Particularly, [16] uses not only all the attributes in our work but also some more attributes such as CH_4, NMHC, CO_2, NO and solar radiation, which means a higher cost. [5] uses another data set including Moderate Resolution Imaging Spectroradiometer (MODIS) data and Meteorological data, so the cost is difficult to be compared with our work.

We also calculate the two criteria reflecting the classification accuracy of discrete and continuous meaning to evaluate our method. As is shown in Fig. 3, the correct points under continuous meaning are in the area between two dot lines. There are 3323 correct results out of 3686 data items, which means the classification accuracy reaches 90.34%. On the other hand, the classification accuracy is only 68.91% under discrete meaning. Lots of mistakes occur near the boundary of an IAQI level because our training is not aimed at the classification of IAQI level.

The application can also be regard as a classification problem of IAQI levels. The data need to be discretized into separate levels according to Table II. The transformation may cause loss of information, so it is actually a weaken process of the problem and model. But the error rate near the boundary between different levels is reduced, which makes it more suitable for classification. The result of our classification method is shown in Table IV. The final classification accuracy increases to 72.80%, which is better than directly using the estimation model to classify.

It can be seen from the result that our method can achieve high accuracy no matter the problem is regard as an estimation or a classification task.

D. Comparison with Other Choices

The comparison result between the final model and other choices, including different input attributes, structures and preprocessing methods, is shown in Table V.

TABLE V
RESULTS OF DIFFERENT STRUCTURE AND INPUT ATTRIBUTE CHOICES

No.	Change	$TrainData$	$TestData$
-	Final Model	0.9284	0.9002
1	With $WindSpeed$	0.9304	0.8991
2	Without CO	0.8788	0.8174
3	Without NO_2	0.9344	0.8714
4	Without O_3	0.8890	0.7661
5	Without PM_{10}	0.8937	0.8811
6	Without SO_2	0.9306	0.8971
7	Without $Temperature$	0.9103	0.8626
8	Without $Humidity$	0.8952	0.8702
9	Without $WindDirect$	0.9278	0.8995
10	500 Hidden Nodes	0.9306	0.8981
11	Linear Ouput Layer	0.8994	0.8634
12	Without Entropy Maximization	0.9396	0.8068
13	Considering Sensor Noise	0.8994	0.8691

After iterative training with the proposed greedy algorithm, we find that the $WindSpeed$ is a redundant attribute which reduces the estimation accuracy. This result is contrary to typical experience because $HighWindSpeed$ that can blow the pollutants away is usually considered a major cause of great air quality. However, we have already used kinds of pollutant data in the model, so the phenomenon that every pollutant data has a low value already contains the information of $HighWindSpeed$. So the $WindSpeed$ attribute has no more effect on $PM_{2.5}$ but further brings noise into the model. As a result, the accuracy increases when we use all other attributes except $WindSpeed$ to estimate $PM_{2.5}$. The result

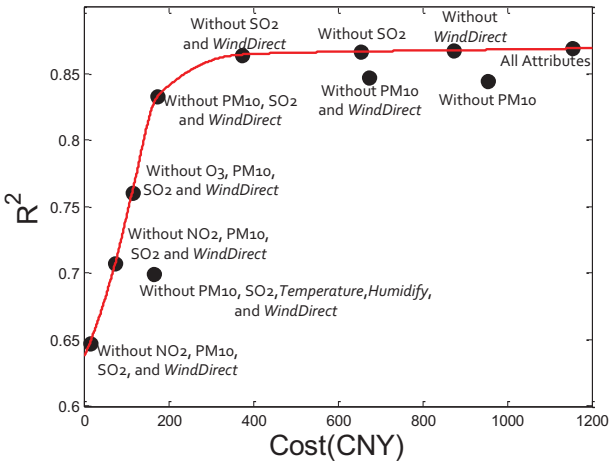

Fig. 4. Relationship between cost and estimation accuracy of our method

of using $WindSpeed$ is shown by comparison No.1 in Table V. And the results of further eliminate another attribute from the optimal input attributes are shown by comparison No.2 to No.9.

We finally use 300 nodes in hidden layer and use sigmoid function in both hidden layer and output layer, which can better fit the multi-layer non-linear model according to the examination results of $ValidationDataset$. The results of using other structures of ANN are shown by comparison No.10 and No.11 in Table V. In addition, it can be seen from the comparison No.12 in Table V that without the $Entropy\ Maximization$ step, the result of $TrainingDataset$ gets better while the result of $TestingDataset$ gets worse. The result means that the $Entropy\ Maximization$ can avoid over-fitting related to the data distribution in $TrainingDataset$ and make the final model more general to other data sets, which matches our expectation.

E. The Cost-Accuracy Relationship

The estimation accuracy decreases when we use fewer attributes than optimal input attributes, while the fewer attributes can lead to lower cost on sensors. In order to further reduce the cost when there is a specific estimation accuracy demand, we create a fit curve to reflect the relationship between cost and accuracy of our method according to the price of each sensor, which is shown in Fig. 4. Each point refers to a choice whose x-coordinate is the cost in Chinese Yuan and y-coordinate is R^2 reflecting the estimation accuracy. In addition, we introduce the typical error rate of each sensor into the experiment, so the linear correlation coefficient R decreases compared with the ideal result. We use the $Additive\ White\ Gaussian\ Noise$ to simulate the error, which is typical when analysing the noise of sensor [17]. The result of noise considered situation is shown by the comparison No.13 in Table V.

V. Conclusion

This work implements an accurate and low-cost method to estimate the concentration of $PM_{2.5}$ based on ANN. We find out two major problems that limit the estimation accuracy, and propose specific algorithms to solve them. First, An $Entropy\ Maximization$ step is proposed to avoid the over-fitting related to data distribution and can make R^2 increase 0.1. Second, we abstract choosing suitable input attributes as an optimization problem and proposed a greedy algorithm to solve it. The result shows that $WindSpeed$ is a redundant attribute and R^2 can increase up to 0.9 after eliminating $WindSpeed$. Additionally, in order to further reduce the cost, we analyse the trade-off relationship between the cost and accuracy using the price and error rate parameters of each sensor. This relationship can help to choice suitable input attributes with specific accuracy demand.

In the future, we will try to predict $PM_{2.5}$ with our method. And we want to find the relationship of $PM_{2.5}$ among different regions.

Acknowledgement

This work was supported by 973 project 2013CB329000, National Science and Technology Major Project 2013ZX03003013-003 and National Natural Science Foundation of China (No.61373026, 61261160501, 61271269), Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation, The Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions, and Tsinghua University Initiative Scientific Research Program.

References

[1] C. Pope et al., "Epidemiology of particle effects," *Air pollution and health*, 1999.
[2] G. Ayers et al., "Teom vs. manual gravimetric methods for determination of pm2. 5 aerosol mass concentrations," *Atmospheric Environment*, 1999.
[3] L. Li *et al.*, "Demonstration abstract: Pimi air box: a cost-effective sensor for participatory indoor quality monitoring," in *Proceedings of the 13th international symposium on Information processing in sensor networks*. IEEE Press, 2014.
[4] Z. Haiming et al., "Study on prediction of atmospheric pm2. 5 based on rbf neural network," in *ICDMA*, 2013.
[5] L. Yao et al., "Ann for multi-source pm2. 5 estimation using surface, modis, and meteorological data," in *iCBEB*, 2012.
[6] Y. Zheng et al., "U-air: When urban air quality inference meets big data," in *SIGKDD*, 2013.
[7] W. J. Parkhurst et al., "Historic pm2. 5/pm10 concentrations in the south-eastern united statesłpotential implications of the revised particulate matter standard," *Journal of the Air & Waste Management Association*, 1999.
[8] W. Zheng, *HJ 633-2012: Translated English of Chinese Standard HJ633-2012: Technical Regulation on Ambient Air Quality Index (on trial)*. www. ChineseStandard. net, 2014.
[9] M. Jarvis *et al.*, "Low cost carbon monoxide monitors in smoking assessment." *Thorax*, 1986.
[10] R. M. Cox, "The use of passive sampling to monitor forest exposure to o3, no2 and so2: a review and some case studies." *Environmental Pollution*, 2003.
[11] S. Kingham *et al.*, "Winter comparison of teom, minivol and dusttrak p-m10 monitors in a woodsmoke environment," *Atmospheric Environment*, 2006.
[12] M.-H. Wang *et al.*, "Using enn-1 to inspect the air pollution of automobile exhaust by remote sensing data," in *Machine Learning and Cybernetics, 2006 International Conference on*. IEEE, 2006.
[13] J. Y. Yang *et al.*, "Effect of choice of kernel in support vector machines on ambient air pollution forecasting," in *ICSSE*, 2011.
[14] W. Haifeng *et al.*, "Research on the assessment for air environment quality based on support vector machine," in *CCDC*, 2009.
[15] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE*, 2001.
[16] S. M. Al-Alawi et al., "Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone," *Environmental Modelling & Software*, 2008.
[17] L. Xiao *et al.*, "A scheme for robust distributed sensor fusion based on average consensus," in *IPSN*, 2005.