An Accurate and Low-cost Method to Estimate $PM_{2.5}$

Lixue Xia, Rong Luo, Bin Zhao, Yu Wang,
Huazhong Yang

Dept. EE, Tsinghua University

Outline

- Background and Motivation
  - Impacts of $PM_{2.5}$
  - Researches about $PM_{2.5}$
- Our Method
  - Use ANN to estimate $PM_{2.5}$
  - Entropy maximization
  - Greedy algorithm to choose input attributes
- Experiment result

The Pollution of $PM_{2.5}$ in China

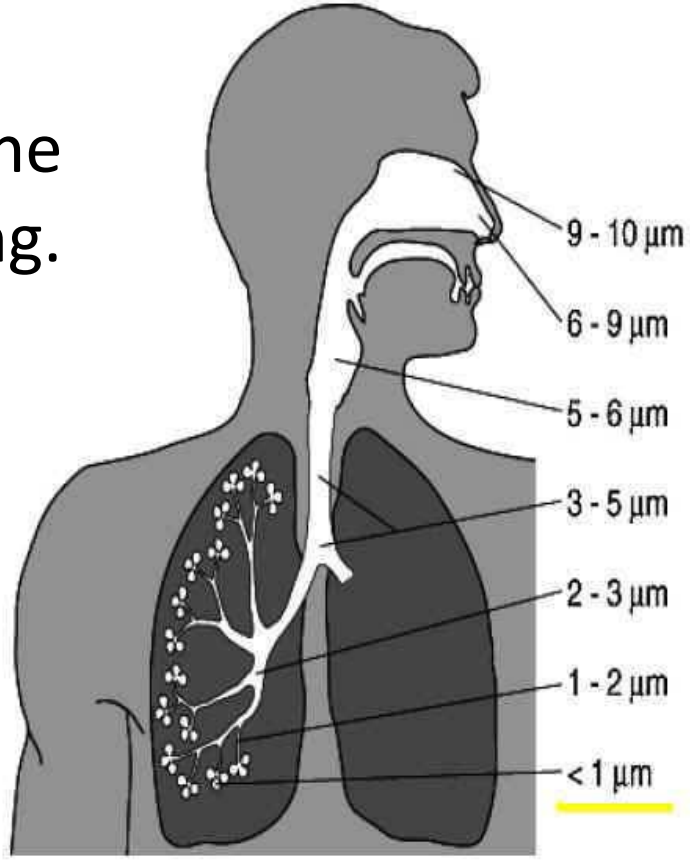
- The increase of $PM_{2.5}$ effects the environment and the human health, which is concerned both by the government and the citizens.

Definition & Health Effect to Human

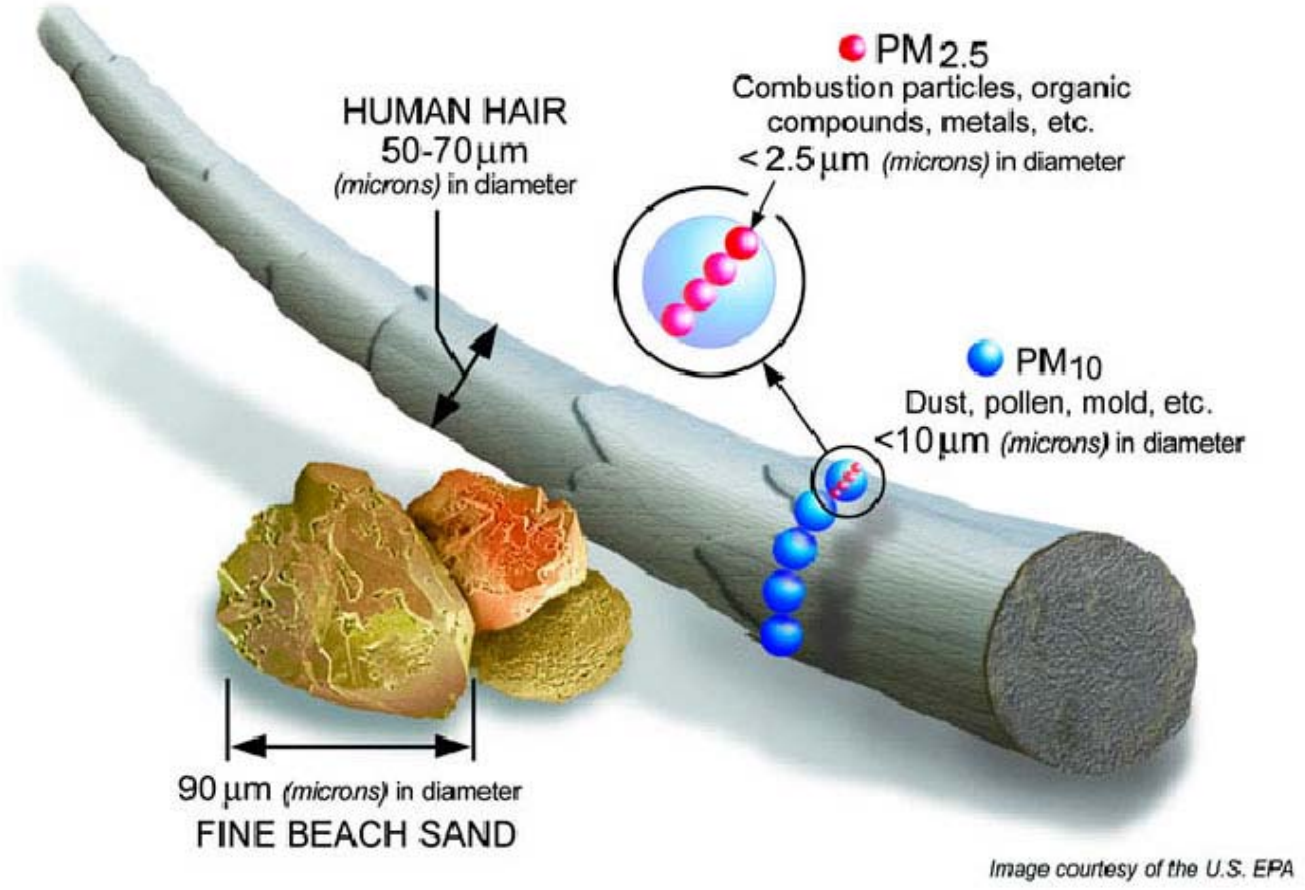
- $PM_{2.5}$, so called fine particle, refers to the Particle Matter whose diameter is less than 2.5 micrometres

- $PM_{2.5}$ tends to penetrate into the gas exchange regions of the lung.

- $PM_{2.5}$ is related to the some serious effects, including lung cancer[1] and other cardiopulmonary mortality.[2]



The Size of $PM_{2.5}$



Researches About $PM_{2.5}$

- $PM_{2.5}$ Measurement or Estimation
  - Measure $PM_{2.5}$ with Physical Method (Light Scattering, Oscillating Microbalance, Spectrum ）
  - Estimate $PM_{2.5}$ by other parameters
- Analysis the cause of $PM_{2.5}$
  - Analysis the formation and deduce the cause
  - Analysis the pollution source and control $PM_{2.5}$
- Analysis the Pathogenic Mechanism of $PM_{2.5}$
  - Analysis the pathogenicity of the component of $PM_{2.5}$
  - Analysis the relationship between illness case and $PM_{2.5}$ concentration

Researches About $PM_{2.5}$

- $PM_{2.5}$ Measurement or Estimation
  - Measure $PM_{2.5}$ with Physical Method (Light Scattering, Oscillating Microbalance, Spectrum）
  - Estimate $PM_{2.5}$ by other parameters
- Analysis the cause of $PM_{2.5}$
  - Analysis the formation and deduce the cause
  - Analysis the pollution source and control $PM_{2.5}$
- Analysis the Pathogenic Mechanism of $PM_{2.5}$
  - Analysis the pathogenicity of the component of $PM_{2.5}$
  - Analysis the relationship between illness case and $PM_{2.5}$ concentration

The Limitation of Existing Measure Method

- Obtain large amounts of measure data is an important base to analysis the cause of $PM_{2.5}$ and to further control the pollution.
  - Due to the complex characteristic of $PM_{2.5}$, it is difficult to directly measure the concentration. All the existing measure equipment use indirect method, which makes them need high cost to reach high accuracy.
  - The high cost on measurement limits the monitoring and the research of $PM_{2.5}$.

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

Outline

- Background and Motivation
  - Impacts of $PM_{2.5}$
  - Researches about $PM_{2.5}$
- Our Method
  - Use ANN to estimate $PM_{2.5}$
  - Entropy maximization
  - Greedy algorithm to choose input attributes
- Experiment result

Use ANN to Estimate $PM_{2.5}$

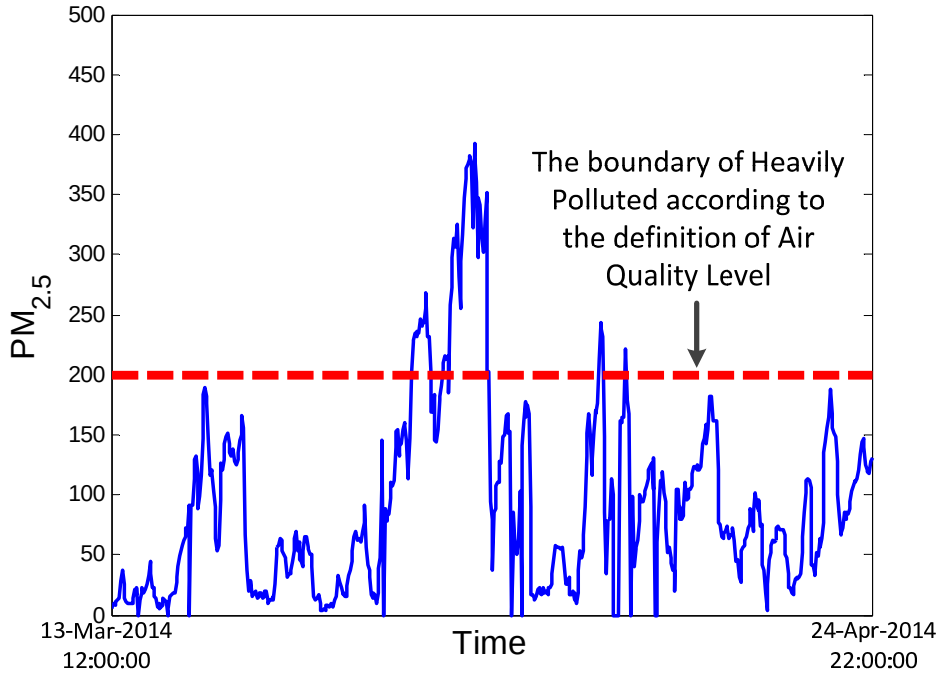
- We use ANN to estimate $PM_{2.5}$ which can be influenced by too many factors compared with other molecular pollutants such as O3 [3].
- The estimation accuracy is low when directly using ANN to estimate $PM_{2.5}$ [4], or the cost goes high again after many kinds of expensive data are used[5].
- We solve two major problems that reduce the estimation accuracy.

Over-fitting on distribution

Redundant input reduces accuracy

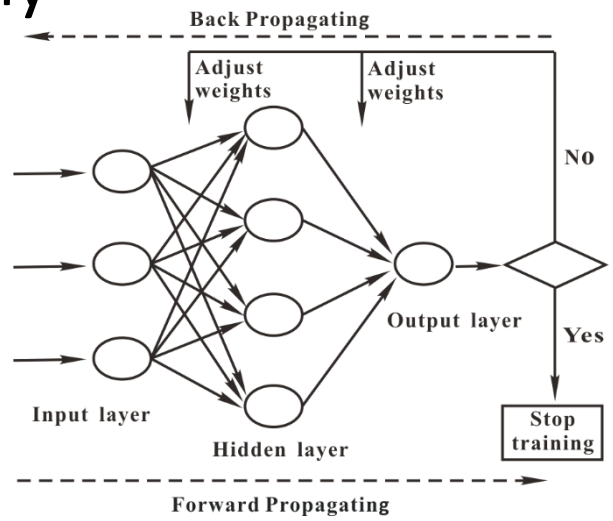
Problem 1 - The Over-fitting on Distribution

- The heavily polluted phenomenon has little probability during time while contains important information.



Rare Data Contains Important Information

- In the training phase, system randomly choose an item to train the model.
  - The range where the event often occur has more data amount, so the NN has more probability to train an item in this range.
  - The low-frequency but high-importance data are despised during the training phase.



- The final trained model has low accuracy on the low-frequency range.
  - This is a kind of over-fitting phenomenon, which is leaded by the non-uniform probability distribution of the $PM_{2.5}$ data.

Entropy Maximization

- *Information Entropy* — Event with lower probability reflects more important information.
- Given a discrete random variate *X* whose probability distribution P = {p_1,p_2,…,p_n}, the entropy of *X* is

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

- The entropy reaches the maximum value if and only if P equals uniform distribution, namely

$$0 \le H(X) \le \sum_{i=1}^{n} \frac{1}{n} \log(n) = log(n)$$

- We should normalize the probability distribution of a variate to the *uniform distribution* to maximize the information entropy and then completely utilize the information.
  - This normalization can also reduce the over-fitting phenomenon on data distribution.

Normalization on Distribution

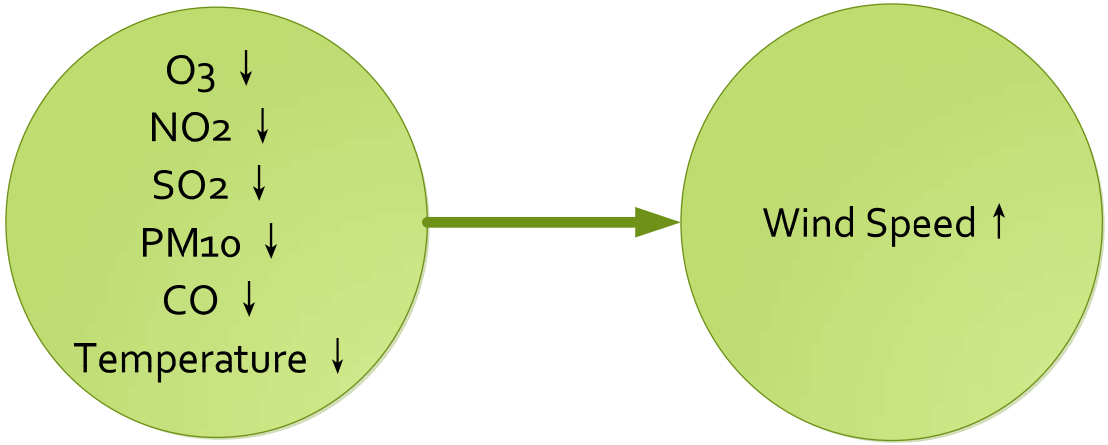
- According to the existing result of probability theory, if we already have a prior distribution knowledge of a random variable *x* whose distribution is p(*x*) with range of interval [0,1], we can normalize *x* into a random variation *y* with uniform distribution using the function:

$$y = g(x) = \int_0^x p(x)dx$$

- We use the distribution of train data as the prior distribution

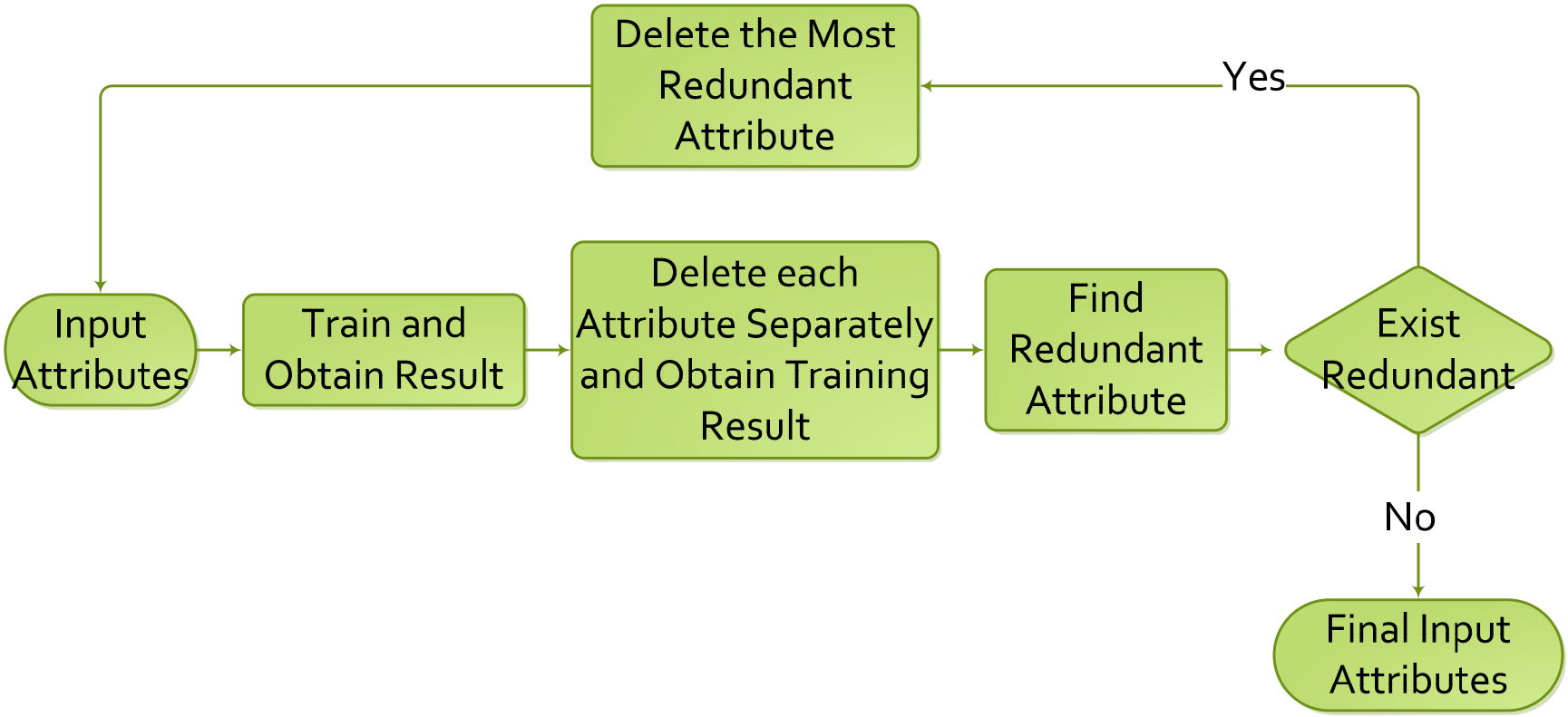
Problem 2 – Redundant Input Reduces Accuracy

- The useful information of some input attributes can be completely reflected by other inputs.
- Given this situation, these redundant inputs can only introduce more noise and interference (i.e. aggregation characteristic over seasons) into the model. The accuracy increases if we delete these redundant inputs from the model.

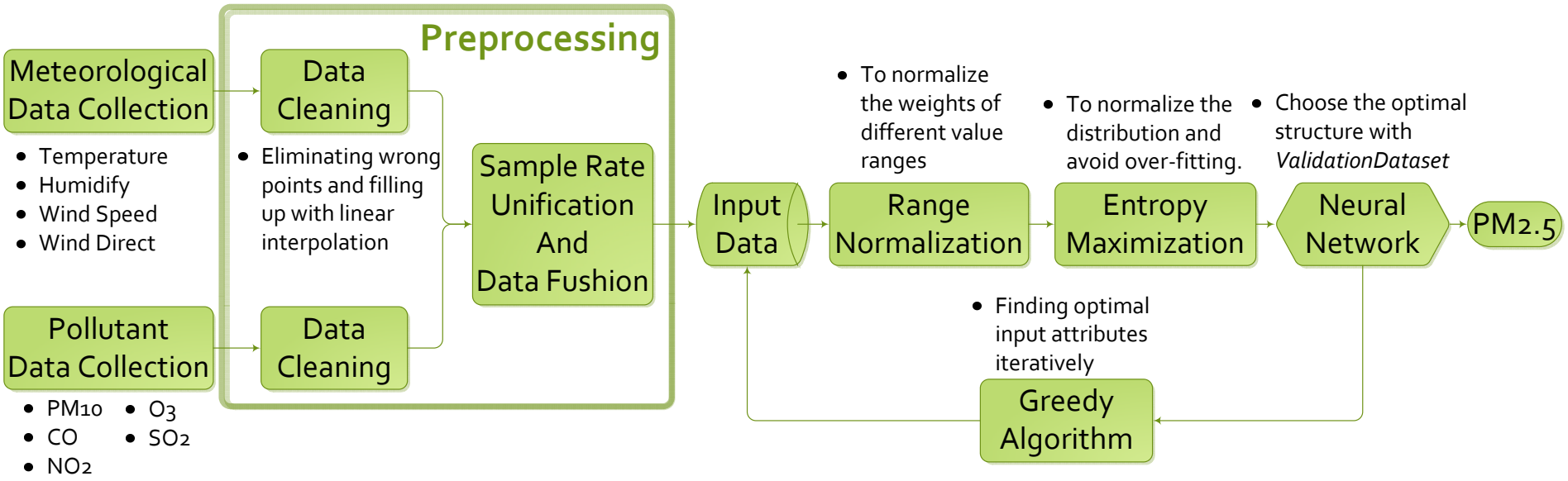


Iterative Greedy Algorithm

- We greedily eliminate the attribute that limits the accuracy most in each iteration and find the optimal input attributes step by step.



Structure of the System



- *TestingDataset* has total 3686 valid data items

Outline

- Background and Motivation
  - Impacts of $PM_{2.5}$
  - Researches about $PM_{2.5}$
- Our Method
  - Use ANN to estimate $PM_{2.5}$
  - Entropy maximization
  - Greedy algorithm to choose input attributes
- Experiment result

Evaluation Criteria

- The linear correlation coefficient *R* between estimated values and real values.
- This problem can also be regarded as a classification of IAQI, we propose 2 kinds of accuracy criteria.
  - The number and proportion of estimated IAQI values classified in the same level with real value.
  - The number and proportion of estimated IAQI values whose difference from the corresponding real value is not larger than 50.

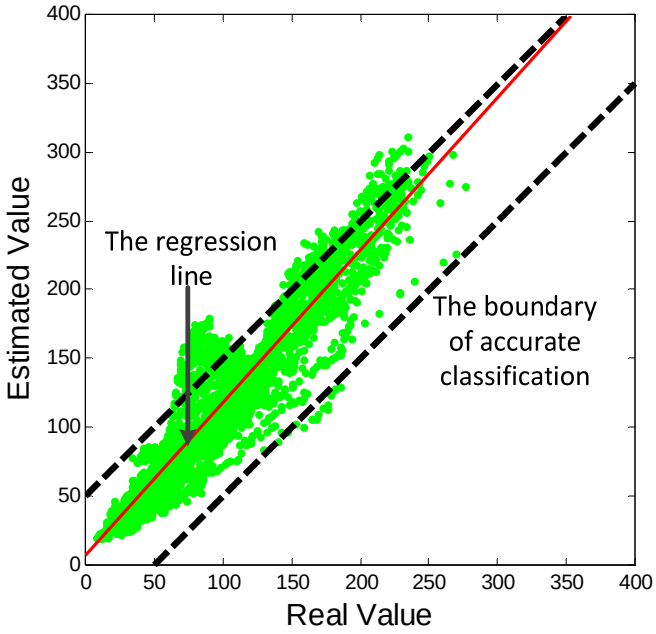
Level	AQI range	Air Quality
1	0-50	Good
2	51-100	Moderate
3	101-150	Lightly Polluted
4	151-200	Moderately Polluted
5	201-300	Heavily Polluted
6	301-500	Severely Polluted

[6]

Experiment Result

- The result of Training Dataset
  $R=0.9488$, $R^2=0.9002$

Works	Pollutant	Data	Data Amount	R^2 with ANN
[3]	O_3	Pollutants & Meteorological data	A week	0.845
[5]	$PM_{2.5}$	Ground monitor, MODIS & Meteorological data	<200 5 days	0.6556
Ours	$PM_{2.5}$	Pollutants & Meteorological data	3686 2 weeks	0.9002



- The classification accuracy reaches 90.34% under continuous meaning.

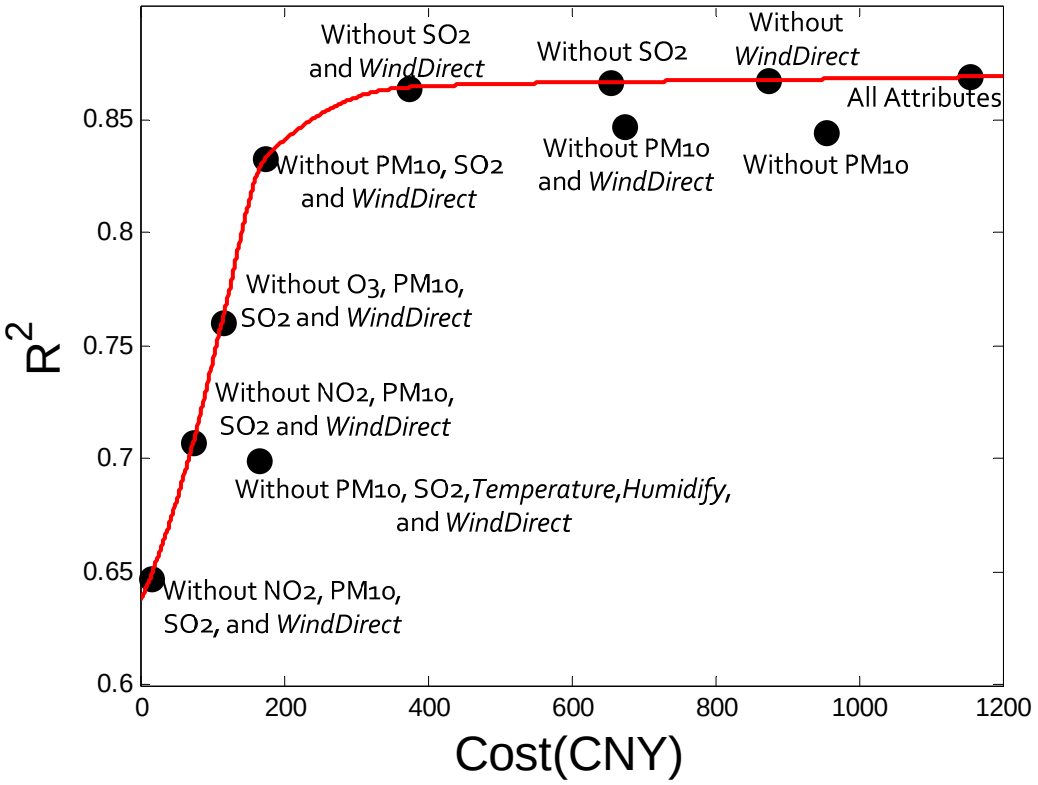
Comparison with Other Choises

No.	Change	$TrainDataset$	$TestDataset$
-	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

- 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.

Trade-off between Cost and Accurancy

- The fewer attributes can lead to lower cost on sensors.
- User can further reduce the cost when there is a specific estimation accuracy demand.



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.
- We analyze 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.

Reference

[1] Raaschou-Nielsen, Ole et al. Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). The Lancet Oncology , Volume 14 , Issue 9 , 813 - 822

[2] Cohen A J, Ross Anderson H, Ostro B, et al. The global burden of disease due to outdoor air pollution[J]. Journal of Toxicology and Environmental Health, Part A, 2005, 68(13-14): 1301-1307.

[3]SalehMAl-Alawi, Sabah A Abdul-Wahab, and Charles S Bakheit. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. Environmental Modelling & Software, 23(4):396–403, 2008.

[4]Wayne Zheng. HJ 633-2012: Translated English of Chinese Standard HJ633-2012: Technical Regulation on Ambient Air Quality Index (on trial). www. ChineseStandard. net, 2014.

[5]Zheng Haiming and Shang Xiaoxiao. Study on prediction of atmospheric pm2. 5 based on rbf neural network. In Digital Manufacturing and Automation (ICDMA), 2013 Fourth International Conference on, pages 1287–1289. IEEE, 2013.

[6]Wayne Zheng. HJ 633-2012: Translated English of Chinese Standard HJ633-2012: Technical Regulation on Ambient Air Quality Index (on trial). www. ChineseStandard. net, 2014.

Thank you !