PSO active learning of XGBoost and spatiotemporal data for PM2.5 sensor calibration

PSO active learning of XGBoost and spatiotemporal data for $PM_{2.5}$ sensor calibration

Peng-Yeng Yin[1], Chih-Chun Tsai and Rong-Fuh Day

Department of Information Management, National Chi Nan University, Nantou, 54561, Taiwan.

[1] Email: pengyengyin@gmail.com

Abstract. Ambient $PM_{2.5}$ concentrations affect human health and natural environment. Government-built $PM_{2.5}$ monitoring supersites are accurate but cannot provide a dense coverage of the air quality index (AQI) monitoring. Broadly-distributed $PM_{2.5}$ microsite sensors can complement supersites for fine-grained monitoring. However, due to the low cost of microsite sensors, the accuracy of the raw AQI measurements is not high enough for monitoring purpose. Calibration of low-cost sensors is thus a necessary processing step to enhance measurement fidelity. This paper presents a particle swarm optimization (PSO) based active learning of optimal configurations of XGBoost and spatiotemporal data for $PM_{2.5}$ microsite sensor calibration. The experimental results show that PSO active learning of the optimal configurations of XGBoost and spatiotemporal data can calibrate low-cost $PM_{2.5}$ microsite sensors to achieve high accuracy by reference to governmental supersites.

1. Introduction

The industrialization and human activities have drastically increased the concentrations of particulate matter with aerodynamic diameter ≤ 2.5 μm ($PM_{2.5}$) in our environment. Many researchers have empirically shown the strong correlation of ambient $PM_{2.5}$ concentrations with human health [1], climate change [2], atmospheric visibility[3], plant species mortality [4], to name a few. $PM_{2.5}$ is a collection of complex compounds and it has multiple sources. The transportation and dispersion path of $PM_{2.5}$ is hard to analyze and predict due to many anthropogenic activities and uncertain scenarios of meteorological conditions such as wind speed and direction, precipitation, temperature, relative humidity, atmospheric pressure, and solar radiation. $PM_{2.5}$ is contributed by natural sources (soils, crustal elements, volcanic eruptions, biomass, etc.) anthropogenic sources (such as vehicle exhaust, coal and gasoline combustion, petrochemical production, and steel refinery), and photochemical transformation of precursor emissions such as SO_2 and NO_x. To realize the complex apportionment of $PM_{2.5}$ concentration, expensive and sparsely-distributed supersite sensors were built by government to monitor possible contaminations at few regions of interest.

To reach a broader coverage of monitoring area, participatory citizens and researchers have built low-cost microsite $PM_{2.5}$ sensors which provide denser but less accurate data than those measured by supersites. A feasible solution to enhance the fidelity of low-cost sensors is to find the relationship function between the measurements of microsite sensors and supersite sensors. The relationship function can be found by multiple linear regression, higher-order regression, support vector regression, gradient regression tree boosting, adaptive neuro-fuzzy inference system (ANFIS), to name a few [5]. Moreover, researches have shown that the geographical landscapes, local land usage and

meteorological patterns have various degrees of influence on $PM_{2.5}$ concentration [6]. We believe that by considering multi-scale data in both spatial and temporal dimensions can improve the accuracy of sensor calibration. This paper proposes a calibration method of low-cost microsite $PM_{2.5}$ sensors by using particle swarm optimization (PSO) [7] for active learning of eXtreme Gradient Boosting (XGBoost) [8] and spatiotemporal data. Our method actively chooses the optimal configurations of XGBoost parameter settings with the best composition of spatial and temporal $PM_{2.5}$ data. The comparative results with supersite sensors show the improving accuracy of the low-cost microsite sensors by using the proposed calibration method.

2. Methodology

XGBoost [8] is a novel gradient tree boosting algorithm which has won several competitions including Kaggle's challenges and KDDCup 2015. By using a sparsity-aware split-finding algorithm and weighted quantile sketch, XGBoost is able to scale up to handle billions of data examples but only consume fewer computational resources than existing machine learning methods. Due to the prestigious performance and data scalability, we apply XGBoost to learn an ensemble of regression trees which best interpret the relationship function between the measurements of microsite sensors and supersite sensors. To reach the best performance with XGBoost, there are a number of algorithmic parameters involved to be optimally tuned. The search range of eligible parameter values and their connotations are listed in Table 1.

Table 1. Value ranges and connotations of XGBoost parameters.

Parameters	Ranges	Connotations
g_1	[1, 4]	Tree maximal level
g_2	[1, 300]	Number of boosting trees
g_3	[0, 12]	Minimum weighted sum of leaf nodes
g_4	[0.001, 0.9]	Learning rate
g_5	[0, 1.0]	Proportion of training data
g_6	[0, 2.0]	Threshold for split finding
g_7	[0, 2.0]	L1 regularization term
g_8	[0, 2.0]	L2 regularization term

To learn the best configuration of XGBoost parameter values with the optimal combination of spatial and temporal data, we apply the particle swarm optimization (PSO) algorithm [7] to accomplish this learning task. PSO is an outstanding evolutionary algorithm which is capable of learning the optimal value of decision variables to an explicit or implicit objective function. PSO has been applied to a wide range of complex domains such as energy demand prediction [9], wind turbine placement [10], educational informatics [11], etc. PSO is a bio-inspired algorithm which mimics the social dynamics of bird flocking or fish schooling. A swarm of birds flock synchronously, change direction suddenly, scatter and regroup iteratively, and finally perch on a position. This form of social intelligence not only increases the success rate for food foraging but also expedites the process. The advantages of PSO include natural metaphor, stochastic move, adaptivity, and positive feedback. The PSO algorithm realizes simple rules and serves as an optimizer for elusive problems.

Now we formally present our learning task as follows. As previously noted, the metric scale of spatial and temporal data should be explored to find the most appropriate composition of training dataset. In this paper, we consider three microsite sensors which are within 200 m to their nearest supersite (see Table 2) for calibration. The choice of spatial training data for each sensor is thus classified into four various-scale categories: using its own measurements ($C = 1$), using measurements of its own and another sensor ($C = 2$ or 3), and using measurements of all three sensors ($C = 4$). The available temporal training data is 60-day historical hourly $PM_{2.5}$ measurements. Considering the

characteristic of time series, we use temporal data in a multi-scale time window in t immediately preceding days for calibration. The search range of t is between 1 and 30.

Table 2. Distance between low-cost and supersite sensors.

Low-cost sensors	Nearest supersite sensors	Distance (m)
A1	B1	122
A2	B2	85
A3	B3	112

In consideration of both XGBoost configuration and the combination of spatiotemporal data, we apply PSO to actively learn the decision task. Figure 1 shows the concept of our active learning approach where the PSO algorithm explores the complex space consisting of possible combinations of models and data. Each model is represented by an instance of XGBoost parameterization of nine variables and each piece of data is a selection of data source from spatial, temporal, or spatiotemporal data composition. Hence, the decision problem results in a mixed integer programming formulation which involves 10 decision continuous or combinatorial variables for indicating a combination instance of XGBoost model (g_1, g_2, …, g_8), chosen category for measurement (C), and number of learning days (t). The representation of the PSO particle is shown in Figure 2.

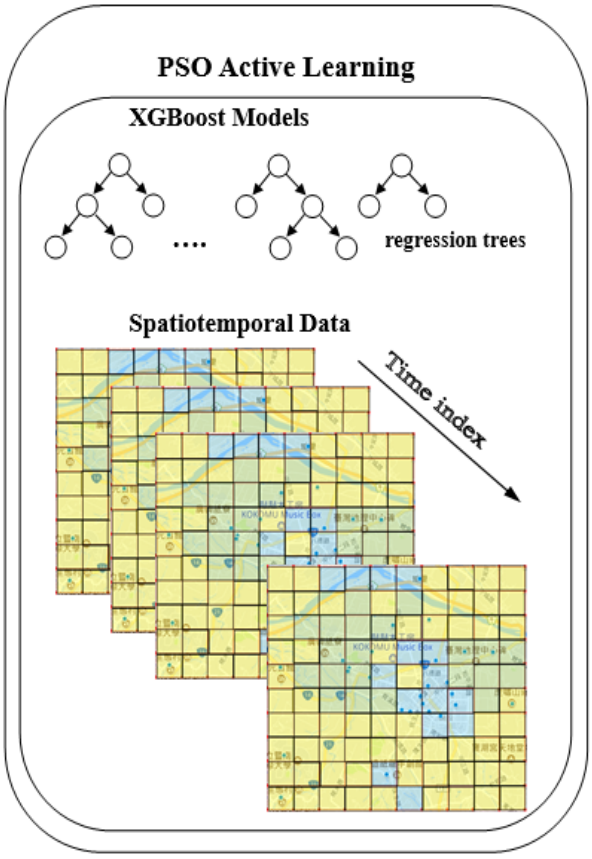

Figure 1. Concept diagram of the proposed approach.

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	C	t

Figure 2. Representation of the PSO particle.

A swarm of particles are employed to explore the enormous model-data space. The rationale of PSO learning is relying on guidance of personal best (*pbest*) and global best (*gbest*) memory identified in the evolution so far by every particle. In particular, particle *i* has a personal memory storing the best position among those it has visited, referred to as $pbest_i$, and the best position *gbest* visited by the entire swarm. The PSO iterates an evolution until a stopping criterion is satisfied, which is usually set as a maximum number of iterations. At each iteration, particle *i* adjusts its velocity v_i and position p_i by referring to the personal best and the swarm's best position as follows.

$$v_i \leftarrow K\left(v_i + c_1 r_1 (pbest_i - p_i) + c_2 r_2 (gbest - p_i)\right) \quad (1)$$

$$p_i \leftarrow p_i + v_i \quad (2)$$

where p_i is the position of the *i*-th particle, *K* is the constriction factor, c_1 and c_2 are the accelerating coefficients, and r_1 and r_2 are random numbers drawn from (0, 1). The fitness (a score of survival) of a particle is defined by evaluating the performance of the model-data combination encoded in the particle with the measure from the referred supersite. We employ three commonly used performance indicators for evaluating the regression models. The three indicators are R-Squared (coefficient of determination), RMSE (root mean squared error), and NME (normalized mean error).

3. Results and discussion

We have deployed several $PM_{2.5}$ microsite sensors around the government-built supersites located in central Taiwan area. The period of collected $PM_{2.5}$ concentration data is between September 24 and November 22 in 2017, in total of 60 days. The data falling in the first 30 days are used for the selection of the used number of days of training data (i.e. t = 1 to 30), while the data of the rest 30 days are used for testing.

We compare the original measures of $PM_{2.5}$ microsite sensors with the calibrated result from four different sources for PSO active learning. Our first calibration method is named PSO-learned XGBoost which deploys PSO to learn the optimal XGBoost parameter values (all g_i). The used spatial and temporal data are fixed to the corresponding microsite (C = 1) and in all 30 training days (t = 30). So there is no active learning with the spatiotemporal data. The second calibration method is named PSO-learned XGBoost and spatial data. The PSO particle includes all g_i and C variable, but fixing t = 30. So this method actively learns the best composition of spatial data from the three microsites. The third calibration method is named PSO-learned XGBoost and temporal data. In this method, the best value of all XGBoost parameters g_i and the time window length variable t are delved. However, the spatial variable is not explored. Finally, the fourth calibration method is named PSO-learned XGBoost and spatiotemporal data uses all g_i, C, and t variables and actively learns the optimal configuration of XGBoost and the best composition of spatiotemporal data.

Table 3 shows the measurement accuracy of the low-cost $PM_{2.5}$ microsite sensors without or with various calibration method. It is seen that the accuracy obtained with all calibration methods is higher than that obtained without calibration, indicating the necessity of calibration process for low-cost microsite sensors. The best performing accuracy for each microsite under various performance metrics are shown in boldface. It is seen that the PSO-learned XGBoost and spatiotemporal data defeats the other three calibration methods by obtaining eight best results out of nine testing cases. So we claim that the best calibrator of all tested methods is the PSO-learned XGBoost and spatiotemporal data. The implication of the finding is significant in the sense that the exploration of complex model-data space is critical to calibration performance, and active learning is a viable approach to this learning task.

Table 3. Calibration residual of microsite sensors.

Calibration methods	R^2	RMSE	NME
Without performing calibration:			
A1	0.7295	12.95	0.5768
A2	0.6555	15.81	0.8877
A3	0.8221	10.83	0.3918
PSO-learned XGBoost:			
A1	**0.7661**	7.38	0.3060
A2	0.6492	7.13	0.3574
A3	0.7997	5.77	0.1743
PSO-learned XGBoost and Spatial Data:			
A1	0.7649	6.99	0.3093
A2	0.6608	8.05	0.3617
A3	0.8058	5.67	0.1749
PSO-learned XGBoost and Temporal Data:			
A1	0.7455	6.62	0.2645
A2	0.6672	6.48	0.3355
A3	0.8267	5.44	0.1714
PSO-learned XGBoost and Spatiotemporal Data:			
A1	0.7625	**6.25**	**0.2627**
A2	**0.6719**	**6.39**	**0.3330**
A3	**0.8296**	**5.12**	**0.1573**

4. Conclusions

In summary, we have proposed a novel calibration approach by using PSO active learning of XGBoost and spatiotemporal data for low-cost $PM_{2.5}$ microsite sensors. By referring to the expensive government supersites, low-cost microsites can be calibrated to provide accurate measures and dense geographic coverage, providing a better surveillance of AQI network. The experimental results show that our method actively learns the optimal configurations of XGBoost model and the best composition of spatial and temporal data to fit the characteristics of $PM_{2.5}$ measurement.

Acknowledgment

This research is partially supported by Ministry of Science and Technology of ROC, under Grant MOST 107-2410-H-260 -015 -MY3.

References

[1] Song C, He J, Wu L, Jin T, Chen X, Li R, Ren P, Zhang L, Mao H 2017 Health burden attributable to ambient PM2.5 in China *Environmental Pollution* **223** 575

[2] IPCC 2007 In: Houghton J T, Ding Y, Griggs D J, Noguer M, Vander Linden P J, Dai X, Maskell K, Johnson C A (Eds.), Climate Change 2007: the Scientific Basis, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University, New York.*

[3] Liu Y J, Zhang T T, Liu Q Y, Zhang R J, Sun Z Q, Zhang M G 2014 Seasonal variation of physical and chemical properties in TSP, PM10 and PM2.5 at a roadside site in Beijing and their influence on atmospheric visibility Aerosol and *Air Quality Research* **14** 954

[4] Mo L, Ma Z, Xu Y, Sun F, Lun X, Liu X, Chen J, Yu X 2015 Assessing the capacity of plant species to accumulate particulate matter in Beijing, China *PLoS One* **10** 0140664.

[5] Ausati S, Amanollahi J 2016 Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting PM2.5 *Atmospheric Environment* **142** 465

[6] Barker H W 2013 Isolating the industrial contribution of PM2.5 in Hamilton and Burlington,

Ontario *Journal of Applied Meteorol. Climatol.* **52** 660
[7] Kennedy J, Eberhart R C 1995 Particle swarm optimization *In Proceedings IEEE Int'l. Conf. on Neural Networks IV 1942*
[8] Chen T, Guestrin C 2016 XGBoost: A Scalable Tree Boosting System In Proceedings of the 22nd ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining (KDD2016) 785 San Francisco, USA*
[9] Askarzadeh A 2014 Comparison of particle swarm optimization and othermetaheuristics on electricity demand estimation: a case study of Iran *Energy* **72** 484
[10] Song M, Chen K, Zhang X, Wang J 2009 Optimization of wind turbine micro-siting for reducing the sensitivity of power generation to wind direction Renewable *Energy* 2016 57
[11] Ho T F, Yin P Y, Hwang G J, Shyu S J, Yean Y N 2009 Multi-objective parallel test-sheet composition using enhanced particle swarm optimization *Educational Technology and Society* **12** 193