Active Learning for Air Quality Station Location Recommendation

S. Deepak Narayanan
IIT Gandhinagar
deepak.narayanan@iitgn.ac.in

Apoorv Agnihotri
IIT Gandhinagar
apoorv.agnihotri@iitgn.ac.in

Nipun Batra
IIT Gandhinagar
nipun.batra@iitgn.ac.in

ACM Reference Format:
S. Deepak Narayanan, Apoorv Agnihotri, and Nipun Batra. 2020. Active Learning for Air Quality Station Location Recommendation. In *7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020), January 5–7, 2020, Hyderabad, India.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3371158.3371208

Motivation: Recent years have seen a decline in air quality across the planet, with studies suggesting that a significant proportion of global population has reduced life expectancy by up to 4 years [1, 2, 5]. To tackle this increasing growth in air pollution and its adverse effects, governments across the world have set up air quality monitoring stations that measure concentrations of various pollutants like NO_2, SO_2 and $PM_{2.5}$, of which $PM_{2.5}$ especially has significant health impact and is used for measuring air quality. One major issue with the deployment of these stations is the massive cost involved. Owing to the high installation and maintenance costs, the spatial resolution of air quality monitoring is generally poor. In this current work, we propose active learning methods to choose the next location to install an air quality monitor, motivated by sparse spatial air quality monitoring and expensive sensing equipment.

Related Work: Previous work has predominantly focused on interpolation and forecasting of air quality [7, 8]. Work on air quality station location recommendation has largely been limited [4]. Previous work [4, 7, 8] has shown that installing air quality stations uniformly to maximize spatial coverage does not work well in practice, which acts as a major motivation for our work.

Problem Statement: Given a set S of air quality monitoring stations, along with their corresponding values of $PM_{2.5}$ over a period of time $\{d_1, d_2, ...d_n\}$, where d_i represents day i, we want to choose a new location s', such that installing a station at s' gives us the best estimate of air quality at unknown locations.

Approach: We perform active learning using Query by Committee (QBC) [6]. We maintain three sets of stations - the train set, the test set, and the pool set. The train set contains currently monitored locations, test set contains the locations where we wish to estimate the air quality and the pool set contains candidate stations for querying, i.e., we query from the pool set and observe how our estimation improves on the test set. To query from the pool set, we need a measure of uncertainty for the stations in the pool set. To obtain this uncertainty, we train an ensemble of learners, and take the standard deviation of their predictions for each station in the pool set. We add the station with maximum standard deviation to our train set, and remove the same station from the pool set. We repeat this process as time progresses. We use K Neighbors Regressor (KNN) as our main model inspired by the fact that nearby days will likely have similar air quality (temporal locality), and so will nearby stations (spatial locality). We create an ensemble of learners by varying the number of neighbors.

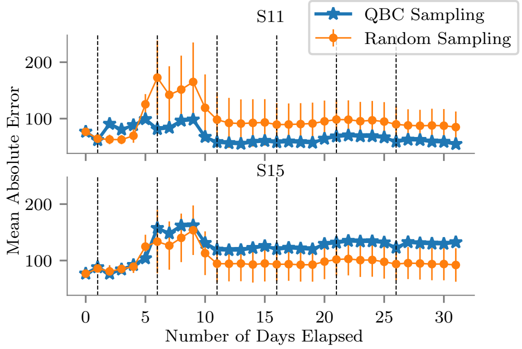

Figure 1: Performance of active learning (QBC) v/s random sampling for sensor installation. The dotted vertical lines denote the days when new stations (from pool set) were added in to the train set of stations.

Evaluation and Results: For the current set of experiments, we use a dataset collected from OpenAQ[1] for New Delhi, for 48 days during the last quarter of the year where there was data consistently available across all the stations. Our feaure set is small: we use only latitude, longitude and time as our features, and we also scale the features between 0 and 1. We have 17 stations in total, and we use 4 stations for training, 1 station for testing and 12 stations for querying. We choose the train stations in a lexicographic manner. We use 15 days worth data for the training set to provide context for the model, and from the 16th day onwards, we query for a station (location) from the pool set every five days. For our KNN model, we create an ensemble of learners by varying the number of neighbors (K) from 1 to 5. As a baseline, we use random sampling, which randomly chooses a station from the pool set for querying. We used 50 different random seeds and present the mean and the standard deviation in Figure 1.

Our main results in Figure 1 shows that QBC active learning performs favourably when compared with random sampling on Station S11 (best performance of QBC) and does only marginally worse than random sampling on Station S15 (worst performance of QBC). We have a win loss ratio of **2.297** and a win percentage of **69.7** for QBC to Random.

Conclusions and future work: We observe that active learning performs better than a random method, using a very small feature set. One natural extension is to use a more extensive feature set for experimentation purposes. Since air quality is affected significantly by factors such as meteorological conditions, traffic conditions, and locations of interests, we could potentially include these factors as features. As an extension to our KNN model, we plan to use Gaussian Processes in future work since they help encode domain knowledge via their ability to support custom kernels [3] and also provide uncertainty along with the predictions.

[1] https://openaq.org

CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371208

REFERENCES

[1] Kalpana Balakrishnan, Sagnik Dey, Tarun Gupta, RS Dhaliwal, Michael Brauer, Aaron J Cohen, Jeffrey D Stanaway, Gufran Beig, Tushar K Joshi, Ashutosh N Aggarwal, et al. 2019. The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017. *The Lancet Planetary Health* 3, 1 (2019), e26–e39.

[2] Yuyu Chen, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. 2013. Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proceedings of the National Academy of Sciences* 110, 32 (2013), 12936–12941.

[3] Vitor Guizilini and Fabio Ramos. 2015. A Nonparametric Online Model for Air Quality Prediction. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, 651–657. http://dl.acm.org/citation.cfm?id=2887007.2887098

[4] Hsun-Ping Hsieh, Shou-De Lin, and Yu Zheng. 2015. Inferring air quality for station location recommendation based on urban big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 437–446.

[5] C Arden Pope III, Majid Ezzati, and Douglas W Dockery. 2009. Fine-particulate air pollution and life expectancy in the United States. *New England Journal of Medicine* 360, 4 (2009), 376–386.

[6] Burr Settles. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison. http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf

[7] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. 2013. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1436–1444.

[8] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. 2015. Forecasting Fine-Grained Air Quality Based on Big Data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 2267–2276. https://doi.org/10.1145/2783258.2788573