Inferring Air Quality for Station Location Recommendation Based on Urban Big Data

Hsun-Ping Hsieh, Shou-De Lin, Yu Zheng

March 29, 2017
Table of Contents

1 Introduction

2 Related Work

3 Data and Features

4 Model and Algorithm

5 Experiments

6 Conclusion
Motivation

- Urban air quality (e.g., concentration of NO₂, PM₂.₅ and PM₁₀,) has attracted more and more attention.
- Air quality index (AQI) is defined to model the pollution levels of the air.
- Accurate air-quality monitoring stations are necessary for AQI measurements.
- However, it is infeasible to construct a lot of monitoring stations due to:
  - space constraint
  - budget constraint
  - labor constraint
- Crowdsourcing based methods are not applicable due to capability constraint on mobile devices.
Objective

- We need a model to recommend locations for monitoring stations.

Problem Definition

Given a set of existing air monitoring stations, where to establish the next ones?
Challenges

- Coverage maximization solution is not applicable since air-quality values are affected by many factors such as weather, traffic, and land usage, which leads to geographically non-smooth values.

- Localizing stations based on inference difficulty is not applicable since we need the ground truth data of all the unobserved locations which is not realistic.

- Localizing stations based on performance improvement maximization is not applicable since we do not really have any observation data about the candidate locations.

- It is difficult to perform an evaluation on the proposed model.
A two-stage framework is proposed.

First stage: create an AQI inference mechanism that not only can infer the AQI values of any arbitrary unobserved location but also reveal the confidence of its inference. A semi-supervised learning framework to infer the air quality values of arbitrary unobserved locations in a city is used.

Second stage: establish new stations at the locations that can minimize the uncertainty of the inference model. A greedy-based entropy-minimization (GEM) is used.
Figure: The proposed framework.
Table of Contents

1 Introduction

2 Related Work

3 Data and Features

4 Model and Algorithm

5 Experiments

6 Conclusion
Inferring Unobserved Sensor Values

- Emission Models. Not applicable due to non-smooth value.
  1. Interpolation models: Inverse Distance Weighting (IDW) and Ordinary Kriging (OK).
  2. Dispersion model.

- Satellite Remote Sensing. Not applicable due to (1) human factors such as traffic and land usage are not considered and (2) sensitivity to weather conditions.

- Crowdsourcing. Not applicable due to (1) sensor capability and (2) sensing time constraint.

- Machine Learning methods. Not applicable based experiment results.
Sensor Deployment Strategies

- Deploying from scratch without observed data.
- Deploying from scratch using observed data. Not applicable due to it does not consider incremental deployment.
- Incremental deployment using observed data.
Table of Contents

1. Introduction
2. Related Work
3. Data and Features
4. Model and Algorithm
5. Experiments
6. Conclusion
Real datasets collected from Beijing air quality monitoring stations is used in this paper.

Air Quality Records. The data contains the real-valued AQI of PM$_{2.5}$ and PM$_{10}$.

Meteorological Data. Five features including temperature, humidity, barometer pressure, wind speed, and weather condition (categorized as cloudy, foggy, rainy, sunny, and snowy) are identified.

Point-Of-Interests (POIs). POI has high correlation to the air quality of the region (e.g. poor air quality might be associated with locations with many factories).

Road Networks. Air quality is strongly affected by the traffic condition.
Table of Contents

1. Introduction
2. Related Work
3. Data and Features
4. Model and Algorithm
5. Experiments
6. Conclusion
Affinity Graph

- We can infer AQI value of one location using information from other locations.
- Using location with station.
- Using near-by locations
- Using recent values
- Using similar layers

Figure: Example of Affinity Graph
Affinity Function

- If two nodes are similar in terms of features, their AQI values are similar to each other.
- For two node $u$ and $v$, feature similarity: $\Delta f_k(u, v) = ||f_k(u) - f_k(v)||$
- Affinity of $u$ and $v$ on one feature $f_k$:
  \[ AF_{f_k}(\Delta f_k(u, v)) = a \cdot \Delta f_k(u, v) + b \]
- For a set of features $F = \{f_1, f_2, \ldots, f_m\}$, affinity of $u$ and $v$:
  \[ a(u, v) = \exp\left(-\sum_{k=1}^{m} \pi_k^2 \times AF_{f_k}(\Delta f_k(u, v))\right) \] which is a softmin of all affinities.
AOI Inference

- AOI distribution of one node $u$: $P(u)$
- Force $P(u)$ to be similar to its close neighbors:
  $$Q(p) = \sum_{(u,v) \in E} w_{u,v} \cdot (P(u) - P(v))^2$$
- Using KL Divergence to measure the difference between $P(u)$ and $P(v)$
- KL Divergence: 
  $$D_{KL}(P(u)||P(v)) = \int_{x=0}^{q_{max}} P(u)[x] \ln\left(\frac{P(u)[x]}{P(v)[x]}\right)$$
- What if $P(v)[x] = 0$?
AQI Inference

- $P(u)$ is a weighted average of its neighbors, which can be better illustrated using example in the Figure. So, what the semi-supervised learning does is to spread knowns AOI values out in a Affinity Graph.
- For mathematical part of proof and derivation, *Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions* and *An overview on the Gaussian Fields and Harmonic Functions Method for Semi-supervised Learning* would be more than helpful.

Figure: Example of Affinity Graph Learning
Minimizing Uncertainty

- Since we have figured out that $P(u)$ can be calculated using weighted average, the weights are only remaining unknown parameters in $Q(P)$.
- Then, the question convert to a optimization problem, which is we want to minimize or maximize something with some constraint on $P(u)$.
- Intuition: maximize the likelihood of labeled nodes using validation data. Suffering data sparsity
- Their method: minimizing the uncertainty of their prediction. The uncertainty can be represented using entropy.
Entropy

- Common form of entropy: \( H(P) = \int_x P(x) \log(P(x)) \, dx \)
- Entropy in this paper:
  \[
  H(P) = \int_x (P(x) \log(P(x)) \, dx + (1 - P(x) \log(1 - P(x))))
  \]
- Objective: minimizing average entropy for all unknown nodes \( U \).
- \( P(u) \) is related to \( w(u, v) \) and
  \[
  w(u, v) = \exp(- \sum_{k=1}^{m} \pi_k^2 \times AF_{f_k}(\Delta f_k(u, v)))
  \]
  which means the unknown thing is \( \pi_k \)
- Using gradient descent method to solve it.
Algorithm

- Extracting features
- Construct affinity graph
- Initialize weights of graph
- Get initial results of $H(P(U))$
- Update $P(U)$ using weights $W$ and then get new $H(P(U))$, then calculated gradient of $H(P(U))$ and update weights $\pi_k$
- Repeat last step until converge.
Algorithm 2

- Identify the location $X_0$ with the lowest entropy
- Choose the most likely value inferred from AQInf and mark $X_0$ as labelled
- Use the pseudo AQI value together with original observed data to build new model
- Identify another location $X_1$ with lowest entropy
- Repeat 1-4 to rank the locations to be recommended from last to first
Table of Contents

1. Introduction
2. Related Work
3. Data and Features
4. Model and Algorithm
5. Experiments
6. Conclusion
Effectiveness of AOInf

Setting
1. Decomposed into 50*50 grids, in which 22 have the monitoring stations containing 10416 time stamps
2. Cross-validation by randomly choosing 15 of 22 grids as labelled data and evaluated by the other 7
3. Repeat 1000 times and average results are reported

3 Models
1. Geographical distances features plus three recent and similar time layers as features (D+T3)
2. Features in (1) plus meteorology data (D+T3+M)
3. Features in (2) plus roadnet and POI features (ALL)

Competitors (3 interpolation-based, 2 learning-based and 2 semi-supervised learning methods)
1. Spatial kNN, Inverse Distance Weighting (IDW), Ordinary Kriging (OK)
2. ANN, SVR
3. Co-training, RBF-SSL
Figure 6. Error of Air Quality Inference at (a) PM$_{10}$ and (b) PM$_{2.5}$ for different algorithms.
Robustness of Air Quality Inference

Figure 7. Error of Air Quality Inference at (a) PM$_{10}$ and (b) PM$_{2.5}$ for different algorithms when R% labeled neighbors removed.
Effectiveness of GEM

Setting
1. Recommend $k=5$ locations to establish monitoring stations
2. Choose 5 locations among 22 to be labelled data and reserve 10 locations to be candidate locations for building new stations, the rest 7 are used for evaluation

Evaluation Metrics
1. Top-rank ratio (TRR), $TRR(C) = \frac{\text{Rank}(C)}{C^\frac{10}{k}}$
2. RMSE-improvement

Competitors
1. Distance-based greedy
2. Temporal feature-dissimilarity greedy
3. Spatial feature-dissimilarity greedy
4. Hybrid feature-dissimilarity greedy
5. Entropy-based search
6. Low edge weight search
Evaluating GEM with Top-rank ratio

Figure 8. The TRR results for PM$_{10}$ and PM$_{2.5}$ with varying number of recommended locations.
Evaluating GEM with RMSE-improvement

Figure 9. The improved RMSE results for PM\(_{10}\) and PM\(_{2.5}\) with varying number of recommended locations.
Evaluating GEM with Entropy variation

Figure 10. The average entropy for PM$_{10}$ and PM$_{2.5}$ with varying number of recommended locations.
Evaluating GEM with different number of training stations

Figure 11. The RMSE of recommending 5 locations for PM$_{10}$ and PM$_{2.5}$ with varying number of labeled locations.
Evaluating GEM with different time spans

Figure 12. The TRR results for PM\textsubscript{10} and PM\textsubscript{2.5} when varying the length of time span $N_T$. 
# Table of Contents

1. Introduction
2. Related Work
3. Data and Features
4. Model and Algorithm
5. Experiments
6. Conclusion
Conclusion

- A model to recommend the most proper locations for air quality monitoring stations is proposed.
- The affinity graph integrates spatial and temporal correlations.
- The weights are learned to not only capture the correlation between features and AQI but also to minimize the uncertainty of the model.
- It is much more effective than myopically minimize entropy or other heuristics.