Forecasting Geo-sensor Data with Participatory Sensing Based on Dropout Neural Network

Jyun-Yu Jiang
Department of Computer Science
University of California, Los Angeles
jyunyu.jiang@gmail.com

Cheng-Te Li
Department of Statistics
National Cheng Kung University
chengte@mail.ncku.edu.tw

ABSTRACT

Nowadays, geosensor data, such as air quality and traffic flow, have become more and more essential in people's daily life. However, installing geosensors or hiring volunteers at every location and every time is so expensive. Some organizations may have only few facilities or limited budget to sense these data. Moreover, people usually tend to know the forecast instead of ongoing observations, but the number of sensors (or volunteers) will be a hurdle to make precise prediction. In this paper, we propose a novel concept to forecast geosensor data with participatory sensing. Given a limited number of sensors or volunteers, participatory sensing assumes each of them can observe and collect data at different locations and at different time. By aggregating these sparse data observations in the past time, we propose a neural network based approach to forecast the future geosensor data in any location of an urban area. The extensive experiments have been conducted with large-scale datasets of the air quality in three cities and the traffic of bike sharing systems in two cities. Experimental results show that our predictive model can precisely forecast the air quality and the bike rentle traffic as geosensor data.

CCS Concepts

•**Information systems** → **Geographic information systems;** •**Mathematics of computing** → *Time series analysis;* •**Computing methodologies** → Neural networks;

Keywords

Geo-sensor data forecasting; Participatory sensing; Urban computing

1. INTRODUCTION

Participatory sensing [6] is a well-known concept to collect various kinds of environmental informatics, such as air quality, traffic, and human mobility. In the typical setting of participatory sensing, the originator pays a set of participants, and asks them to carry some geosensors (e.g. mobile devices or air-quality sensors) and move in the targeted geographical area within a certain time period. To collect sufficient data in terms of geography and time, the participants are required to avoid either staying at a location too long or moving to other places that had been sensored by other participants. The government or some organizations can take advantage of geosenor data to predict future environmental informatics.

However, the environmental informatics may change rapidly, and nearby locations could have significantly different geosensor values. For example, air quality highly depends on the weather condition and human behaviors. The air quality may be quite different in the morning and noon. Moreover, a location with unhealthy air quality may be only one mile away from a healthy place [11]. Hence, forecasting environmental informatics based on the geosensor data collected from the paradigm of participatory sensing may lead to unsatisfying results. In addition, while the targeted area to be sensed is too large, it is unrealistic to assume that we can recruit a large number of participants. Therefore, the geosensor data collected from participatory sensing is supposed to be very sparse in both spatial and temporal aspects.

In this paper, we attempt to solve the data sparsity problem of participatory sensing for forecasting environmental informatics. Given a small set of participants carrying sensors, based on the concept of participatory sensing, we assume that they randomly appear at different locations at different time point to collect geosensor data. Our goal is to exploit such sparse and incomplete geosensor data to forecast the future environmental informatics (i.e., the air quality and the rental traffic of bike sharing in this paper) for all the locations in the targeted geo-spatial area. We propose a neural network-based approach with several useful features, including statistical, temporal and spatial features. However, participatory sensing may lead to several unsensed features. To solve this problem, we present an approach inspired by *dropout neural network* [7], which is a method to prevent overfitting. In the end, we also conduct extensive experiments to demonstrate that our approach can precisely forecast geosensor data with sparse and incomplete geosensor data.

In the literature, there are two relevant studies [3,12]. The work [12] assumes that the air quality values of all the monitoring stations are known, and uses such *complete* geosensor data to predict the future air quality. The other work [3] aims at inferring the *past* air quality values for arbitrary locations using *complete* geosensor data. It is apparent that our forcasting based on participatory sensing is more challenging due to spatio-temporal data sparsity. In addition,

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983902

(a) Traditional Approach (b) Participatory Sensing

Figure 1: The scenarios of forecasting geosensor data with traditional approach and participatory sensing.

our setting is more realistic and general since not all kinds of environmental informatics can have monitoring stations to obtain observed data.

2. PROBLEM STATEMENT

In this section, we introduce the problem statement of forecasting geosensor data with participatory sensing.

Motivation. Figure 1 illustrates the forecasting of geosensor data based on traditional approach and participatory sensing. To forecast geosensor data, most of previous work tend to set up a sensor at every location so that the model can have sufficient historical data monitored to predict the future data. However, it is too expensive to afford sensors at all locations. Furthermore, setting up additional sensors for new locations will be an extra cost. Hence, we propose to forecast geosensor data with participatory sensing.

Consider that we have hired k participants carrying sensors (with a limited budget), at every time point, each of them will randomly bound for a location (which is different from destinations of other participants) and utilize the sensor to record the geosensor data. Through participatory sensing, we can have geosensor data of k locations at every time point. The goal of this paper is to aggregate the data from participatory sensing and contextual information (e.g. meteorology data) to predict the future geosensor data.

Problem Definition. We formally define the problem of forecasting geosensor data with participatory sensing. Let C be the set of locations for forecasting. For each location $c \in C$, $L(c) = (lo(c), la(c))$ shows its longitude and latitude. The geosensor data of location $c \in C$ at the time $t \in \mathbb{Z}$ is denoted as $r(c,t)$, where each time point represents a one-hour duration. $M(c,t)$ represents the meteorology information of the location $c \in C$ at the time $t \in \mathbb{Z}$, including weather, temperature, pressure, humidity, wind speed and direction. Given the number of participants (or the budget) k ($1 \leq k \leq |C|$), $S(t)$ denotes the set of k observed locations. For each location c and each time t, our goal is to predict the geosensor data $r(c,t)$ with the observed data $\{r(c',t') \mid c' \in S(t'), t' < t\}$ and meteorology information $\{M(c',t') \mid c' \in S(t'), t' < t\}$.

3. PROPOSED APPROACH

In this section, we propose a neural network based approach with various useful features. To handle some missing features due to participatory sensing, the concept of dropout neural network is borrowed to solve the problem. We also conducted the theoretical analysis to show the reasonableness of our approach.

3.1 Feature Extraction

To train the model, we extract several useful features from the observed data and meteorology information.

3.1.1 *Statistical Features*

The statistical information, derived from the historical data, is important to understand the knowledge of the targeted informatics Hence, we extract some statistical features from the observed data.

Overall Statistics. The geosensor data of a location may be consistent, so the statistics of past observations should be meaningful. Here we calculate the mean and median of data observed at the target location c before time t as two numerical features.

Hour-based Statistics. Although the data of a location may be consistent in day-level, there still may be dissimilar patterns for different hours. Hence, we compute the statistical values of observed data at the target location c by 24 hours as 2 numerical features.

3.1.2 *Temporal Features*

To predict time-series data, temporal information is much more important. Therefore, features about time and data observed in past are extracted as the temporal features.

Current Time. The time t itself may be important since it can represent the stage of a day or a year. Intuitively the time stage can also be helpful to make the prediction more precise. Therefore, the hour of t is treated as a 24-dimensional categorical feature.

Past Observed Data. The geosensor data may be dependent upon historical data. Many previous work [1,8,9,12] also consider the past sensed data as time series data, such as forecast air quality and ozone level. Here we take the data observed at the target location c in past 48 hours $\{r(c,t-i) | 1 \leq i \leq 48, c \in S(t-i)\}$ as 48 numerical features. Note that there will be several missing features because of participatory sensing. The solution to handle this problem will be introduced in Sec. 3.2.

3.1.3 *Spatial Features*

The geosensor data are all correlated with a geographical location, so the spatial information may be essential to forecast the future data.

Neighbor Observed Data. The geosensor data may be similar at close locations. In other words, the data at near locations may be a good indicator for geosensor data forecasting. Furthermore, the data from near locations can to some degree solve the problem of unsensed features.

To encode the information from other locations, we group

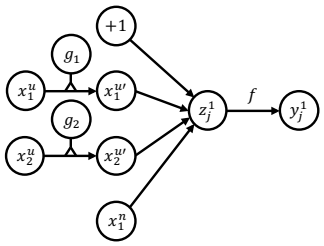


Figure 2: The illustration of our solution inspired by dropout neural network to handle the unsensed features. Note that x_1^u and x_2^u are features possible to be unsensed, and x_1^n is the feature always available. z_j^1 denotes the input into the first hidden layer, and f is the activation function.

them into 24 neighbor groups G_i^c by the distance (3 levels) and the bearing (8 levels) to the location c. For each group G_i^c, the median of observed data at locations in the group at the last time point $\{r(c', t-1) \mid c' \in G_i^c, c' \in S(t-1)\}$ is calculated as a numerical feature. Hence, the geosensor data around the target location can be encoded in the model. Another benefit of grouping locations is that there may be multiple locations in each group. Compared to observing each location separately, the probability of missing features will be lower from p to $p^{|G_i^c|}$, where p is the missing rate of a single location $1 - k/|C|$.

Meteorology Information. The meteorology information is also an important factor reflecting the geosensor data. For example, temperature can significantly affects air quality. Wind directions affect the transportation of air pollution materials. Here we apply the meteorology data $M(c', t-1)$ of the target location c and neighbor groups G_i^c at time $t-1$ as features. For numerical information like temperature and wind speed, we encode them as numerical features, and calculate the median for each neighbor group. For discrete information like weather and wind directions, we encode them as categorical features, and compute the mode for each neighbor group. In this paper, we adopt weather, temperature, pressure, humidity, wind speed and direction as the meteorology information, including $4 \times (1 + 24)$ numerical features and $2 \times (1 + 24)$ categorical features.

Note that even though gathering locations into neighboring groups can reduce the missing features, there may be still several spatial features unsensed by participatory sensing. Moreover, the missing rate of meteorology data at the target location c may not be reduced. Therefore, we also apply the solution mentioned in Sec. 3.2 to solve this problem.

3.2 Training with Unsensed Features

In this paper, we utilize the artificial neural network to train a model with features mentioned in Sec. 3.1. Note that the number of hidden layers can be any number. We simply assign one hidden layer for evaluating the idea.

As mentioned in Sec. 3.1, a problem of participatory sensing is that some features will be unsensed and missing. However, the case in participatory sensing is different from other scenarios of missing data. Because the participants carry sensors choose the locations randomly, the distribution of missing should also correspond with the same distribution. In other words, from the other aspect, the phenomenon of unsensed features can be treated as the results that we randomly ignore or remove them. Dropout [7] is an approach to randomly drop units from neural networks and thereby dealing with overfitting. Inspired by dropout, we consider that the input units of unsensed features are dropped during training. Figure 2 further demonstrates the illustration of our solution. We first divide features into two sets, including a set X^u that might be unsensed and X^n that might not. For each feature $x_i^u \in X^u$, we set a flag g_i to record whether x_i^u is sensed. Based on the number of participants and the number of locations in the neighbor groups, the probability of being sensed can be computed as p_i for the feature x_i. Therefore, from the viewpoint of all data points, the input into the first hidden layer can be treated as follows:

$$z_j^1 = \sum_{x_i^u \in X^u} (x_i^u \cdot g_i \cdot w_i^u) + \sum_{x_i^n \in X^n} (x_i^n \cdot w_i^n) + b,$$

where $g_i \sim \text{Bernoulli}(p_i)$, and b is the bias term. Through this approach, these unsensed features will no longer be a defect. The model can utilize them to prevent overfitting. Finally, we can learn an artificial neural network model with unsensed features.

4. EXPERIMENTS

Datasets. We conduct experiments on five datasets to forecast the air quality (AQ) [11, 12] and the rentle traffic of bike sharing (BS) [5, 10] respectively. For the AQ task, datasets covering three major Chinese cities, including Beijing (BJ), Tianjin (TJ) and Guangzhou (GZ), are applied. The AQ task aims to forecast the $PM_{2.5}$, which is an important air quality index (AQI). Note that although some previous work utilize data from stations near the target city as additional information, we apply only the stations in the city because participatory sensing is based on a limited budget. For the BS task, the experiments are conducted on datasets of two US cities, including Washington D.C. (DC) and New York City (NYC). The BS task is to forecast the rentle traffic in a bike sharing system, which is the number of users who rentle bikes in each time duration. Datasets in both tasks contain the meteorology information. For hours with multiple meteorology records, we compute the values of median and mode for aggregating numerical and categorical information. Table 1 further provides detailed data statistics.

Table 1: Data statistics of five datasets in two tasks.

Dataset	Air Quality (AQ)			Bike Sharing (BS)	
	BJ	TJ	GZ	DC	NYC
Time span	2014/05/01-2015/04/30			2014/04/01-2014/04/30	
# Stations	36	27	42	351	325
# Records	278,085	191,167	283,735	5,359,995	1,886,144
# Meteorology Records	116,867	106,614	30,305	683	449

Experimental Settings. For each dataset, every station is treated as a location to sense so that the ground truth can be guaranteed. We equally divide the time span of each dataset as training and testing periods for evaluation. For every time t, k locations are randomly sampled as the set $S(t)$. Basically, we set k as 50% of locations in each dataset. For analysis, we also evaluate the performance of the BJ dataset with different k.

Table 2: The overall performance of five methods in five datasets for air quality (AQ) and bike sharing (BS) tasks. Note that k is set as 50% of locations. All improvements of our method are significant differences at 95% level in a paired t-test.

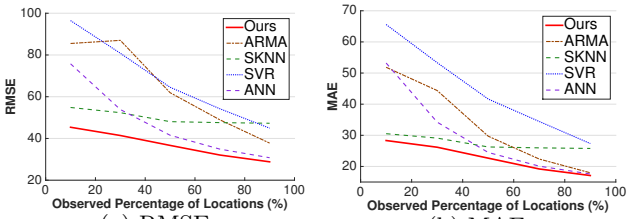
Dataset		Metric	ARMA	SKNN	SVR	ANN	Ours
AQ	BJ	RMSE	61.9305	48.0565	64.5315	41.5409	**36.6724**
		MAE	29.7170	26.2598	41.6494	24.4895	**22.6632**
	TJ	RMSE	49.3223	34.8156	54.7269	37.1958	**31.6178**
		MAE	26.1218	20.8746	37.1335	23.5406	**20.6767**
	GZ	RMSE	17.5112	16.1787	23.7665	13.8771	**13.4062**
		MAE	9.4470	10.6069	16.5019	9.3448	**9.2335**
BS	DC	RMSE	2.8287	2.6056	2.0471	2.0138	**1.8594**
		MAE	1.2328	1.4400	0.9726	0.9546	**0.8955**
	NYC	RMSE	5.9696	4.9828	4.4075	4.1515	**3.8225**
		MAE	3.4115	2.9644	2.5854	2.4919	**2.1987**

Competitive Baselines. Our approach is compared with four baselines: (1) Auto-Regression-Moving-Average (ARMA): ARMA [2] is a well-known model for time-series data prediction. It predicts the geosensor data only by previously sensed data. For the data unsensed by participatory sensing, we directly ignore them and concatenate the sensed data as the input of ARMA. (2) Spatial k-Nearest Neighbors (SKNN): SKNN forecasts the geosensor data by aggregating the sensed data of k spatially closest locations at time $t-1$. (3) Support Vector Regression (SVR): SVR is a variation of support vector machine (SVM) for regression. We utilize the past observed data as features, and set unsensed features as 0. (4) Artificial Neural Network (ANN): We also feed only the past observed data into ANN. Note that it is applied to our approach mentioned in Sec. 3.2 as well.

Experimental results. Root-mean-square error (RMSE) and mean absolute error (MAE) [4] are adopted as the evaluation metrics in our experiments. Table 2 shows the overall performance of five methods. For the AQ task, SVR and ARMA perform the worst among the baseline methods. The reason might be the disappearance and discontinuity of data. Conversely, SKNN performs better because there are always k locations sampled to be the neighbors. ANN performs the best. It also indicates that our approach to handle unsensed features is effective. For the BS task, SKNN performs worse because the rentle traffic may not be relevant to neighbor locations. Besides, ANN still achieves the best performance among the baseline methods. Our approach outperforms all baseline methods in all datasets. It shows that the features in the model and the method to handle unsensed features are so effective. Figure 3 shows that the performance in the BJ dataset with different observed percentages (i.e., the number of participants or sensed locations). It is obvious that all methods perform better when the percentage increases. In addition, our approach performs best among all methods. We also evaluate our approach with different number of the hidden layer as shown in Table 3. With more neurons in the hidden layer, the model is able to achieve better performance.

5. CONCLUSIONS

In this paper, we propose a novel idea of forecasting geosensor data with participatory sensing. Through a small set of participants carrying sensors, we attempt to exploit the sparse and incomplete data to predict the future environmental informatics. A neural network-based solution with features in several aspects is proposed to solve the problem. We also provide a concept inspired by dropout neural network to handle the problem of unsensed features. Experiments conducted on five real-world datasets exhibit the promising effectiveness of our approach. The ongoing future work is to consider the sequential properties of time series data into the model.



(a) RMSE (b) MAE

Figure 3: The performance in the Beijing (BJ) dataset of the air quality (AQ) task with different observed percentage of locations (i.e., the number of participants).

Table 3: The performance of our approach in the Beijing (BJ) dataset of the air quality (AQ) task with different sizes of the hidden layer.

Size	32	64	128	256	512	1024
RMSE	42.5288	39.5373	37.7825	36.6724	36.0191	35.7773
MAE	27.3475	25.1453	23.4175	22.6632	21.8336	20.9201

6. ACKNOWLEDGMENTS

This work was sponsored by Ministry of Science and Technology of Taiwan under grant 104-2221-E-001-027-MY2.

7. REFERENCES

[1] T. Dye, C. MacDonald, and C. Anderson. Guideline for developing an ozone forecasting program. Technical report, Environmental Protection Agency, 1999.
[2] E. J. Hannan. *Multiple time series*, volume 38. John Wiley & Sons, 2009.
[3] H.-P. Hsieh, S.-D. Lin, and Y. Zheng. Inferring air quality for station location recommendation based on urban big data. In *KDD '15*, pages 437–446. ACM, 2015.
[4] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.
[5] Y. Li, Y. Zheng, H. Zhang, and L. Chen. Traffic prediction in a bike-sharing system. In *SIGSPATIAL GIS '15*. ACM, 2015.
[6] F. Restuccia, S. K. Das, and J. Payton. Incentive mechanisms for participatory sensing: Survey and research challenges. *TOSN*, 12(2):13:1–13:40, 2016.
[7] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.
[8] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov. Real-time air quality forecasting, part i: History, techniques, and current status. *Atmospheric Environment*, 60:632–655, 2012.
[9] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov. Real-time air quality forecasting, part ii: State of the science, current research needs, and future prospects. *Atmospheric Environment*, 60:656–676, 2012.
[10] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *TIST*, 5(3):38, 2014.
[11] Y. Zheng, F. Liu, and H.-P. Hsieh. U-air: When urban air quality inference meets big data. In *KDD '13*, pages 1436–1444. ACM, 2013.
[12] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li. Forecasting fine-grained air quality based on big data. In *KDD '15*, pages 2267–2276. ACM, 2015.