Crowdsourcing-based Urban Anomaly Prediction System for Smart Cities

Chao Huang, Xian Wu and Dong Wang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
chuang7@nd.edu, xwu9@nd.edu, dwang5@nd.edu

ABSTRACT

Crowdsourcing has become an emerging data collection paradigm for smart city applications. A new category of crowdsourcing-based urban anomaly reporting systems have been developed to enable pervasive and real-time reporting of anomalies in cities (e.g., noise, illegal use of public facilities, urban infrastructure malfunctions). An interesting challenge in these applications is how to accurately predict an anomaly in a given region of the city before it happens. Prior works have made significant progress in anomaly detection. However, they can only *detect* anomalies after they happen, which may lead to significant information delay and lack of preparedness to handle the anomalies in an efficient way. In this paper, we develop a *Crowdsourcing-based Urban Anomaly Prediction Scheme (CUAPS)* to accurately *predict* the anomalies of a city by exploring both spatial and temporal information embedded in the crowdsourcing data. We evaluated the performance of our scheme and compared it to the state-of-the-art baselines using four real world datasets collected from 311 service in the city of New York. The results showed that our scheme can predict different categories of anomalies in a city more accurately than the baselines.

Keywords

Crowdsourcing, Anomaly Prediction, Bayesian Inference, Smart Cities

1. INTRODUCTION

This paper presents a new Crowdsourcing-based Urban Anomaly Prediction Scheme (CUAPS) to address the urban anomaly prediction problem for smart cities. Urban environment (e.g., air, noise, river) and infrastructures (e.g., buildings, roads, parking lots) are essential parts of a well-functioning city and monitoring their conditions has recently received a significant amount of attentions [3, 9–11]. With the ubiquity of Internet connectivity and the proliferation of smartphones, crowdsourcing has become an emerging data collection paradigm for smart city applications (e.g., location-based services, environment monitoring, information distillation) [3–6]. Along this trend, a new category of crowdsourcing-based urban anomaly reporting systems have been developed to allow common citizens to report, track, and comment on the urban anomalies they encountered in their daily lives (e.g., noise, illegal use of public facilities, urban infrastructure malfunctions). An interesting challenge in these applications is how to accurately predict an anomaly in a given region of the city before it happens. Accurate prediction of anomalies can help the governments and communities to effectively prevent anomalies from happening or handle them efficiently if they happen [7].

Several technical challenges exist in order to solve the anomaly prediction problem by exploring the crowdsourcing data from common citizens in a city. First, *Sparse Crowdsourcing Data*: the crowdsourcing reports are often sparse and incomplete since people may not report anomalies all the time at all places in a city. Furthermore, some of their reports miss important information (e.g., time, location, description of the anomaly) and hence are less useful. Thus, it is difficult to predict the anomaly of a given region from the crowdsourcing data from that region alone. Second, *Uncertain Geographic Dynamics*: anomalies in different geographic regions of the city often have very different probability distributions (e.g., the distribution of noise reports in a downtown subdivision is likely to be different from the one of a quiet subdivision in the suburb). Therefore, it is also not a trivial task to use crowdsourcing reports from one region of the city to predict the anomaly of another.

To address the above challenges, we develop a new Crowdsourcing-based Urban Anomaly Prediction Scheme (CUAPS) to predict the anomaly in a given region of a city by exploring both spatial and temporal information embedded in the crowdsourcing data. In particular, we first develop a Bayesian inference model to identify dependent regions in terms of their anomaly distributions. Then we derive an optimal abnormal state prediction scheme that predicts the anomaly in a region from both its own historical data using a Markov model and the data from its dependent regions. Finally, we evaluate the CUAPS approach using four real-world 311 service datasets collected from the city of New York. The evaluation results showed that our scheme can predict different categories of anomalies in a city more accurately than the state-of-the-art baselines.

In summary, our contributions are as follows:

- This paper addresses the problem of urban anomaly prediction using crowdsourcing data that is publicly available. (Section 2)

CIKM '16, October 24–28, 2016, Indianapolis, IN, USA.

DOI: http://dx.doi.org/10.1145/2983323.2983886

- We develop a principled framework (i.e., CUAPS scheme) that allows us to derive an *optimal* solution to accurately predict the anomaly of each region in a city by exploring the dependency of anomaly occurrence between different regions and the historical anomaly distribution of an individual region. (Section 3)

- We perform experiments to compare the performance of our CUAPS scheme with the-sate-of-the-art baselines using real-world data sets collected from 311 service in New York. Experimental results demonstrate that the proposed approach outperforms existing methods in terms of prediction accuracy. (Section 4)

2. PROBLEM FORMULATION

In this section, we formulate the problem of urban anomaly prediction using crowdsourcing data. In particular, we consider a set of X geographic regions of a city, which are indexed by $i = 1, ..., X$ and each region has Y_i anomalies till time slot K (i.e., the time slot before the predicted slot). We further define the following inputs to our model.

- *Definition 1.* **Anomaly Trajectory**. A trajectory Tr_i of region i is an anomaly trace reported by the crowd in chronological order, e.g., $Tr = l_1 \rightarrow ... \rightarrow l_z \rightarrow ... \rightarrow l_Z$, where each reported anomaly consists of a geospatial coordinates and a timestamp, i.e., $l_z = (latitude, longitude, timestamp)$.

- *Definition 2.* **Abnormal State Vector**. We define the abnormal state vector $AS_{X\times K}$ to represent if there exists reported anomaly of each region in K time slots. In particular, $AS_{i,k} = 1$ denotes that there exists a reported anomaly in kth time slot and $AS_{i,k} = 0$ otherwise.

- *Definition 3.* **Temporal Vector**. For each region, we define the temporal vector t_{Δ_i} as the time difference of consecutive anomalies in that region. This feature has also been considered in [2].

For each region, we consider two features (i.e., the number of anomalies Y_i and temporal vector t_{Δ_i}) in a Bayesian inference model. In our model, the anomaly distribution of each region belongs to one of C clusters (which are indexed by $c = 1, ...C$). Each cluster consists of regions with similar anomaly distributions. We assumes the Multinomial distribution for each cluster and use a common Dirichlet prior to accommodate the possible difference in anomaly distributions between regions within the same cluster.

The problem of inferring the urban anomaly prediction using crowdsourcing data is formulated as follows: given the anomaly trajectory Tr_i till time slot K ($Tr_{i,K}$) of each region in a city, the goal is to predict the unknown abnormal state of all regions in the next time slot $K+1$. Formally, we compute:

$$\forall i, 1 \leq i \leq X : P(AS_{i,(K+1)}|Tr_{i,K}) \quad (1)$$

3. CROWDSOURCING-BASED URBAN ANOMALY PREDICTION

The developed CUAPS consists of three components: Clustering Dependent Regions, Markov Trajectory Estimation and Optimal Anomaly Prediction. We will discuss each component in detail in this section. Figure 1 shows an overview of the CUAPS.

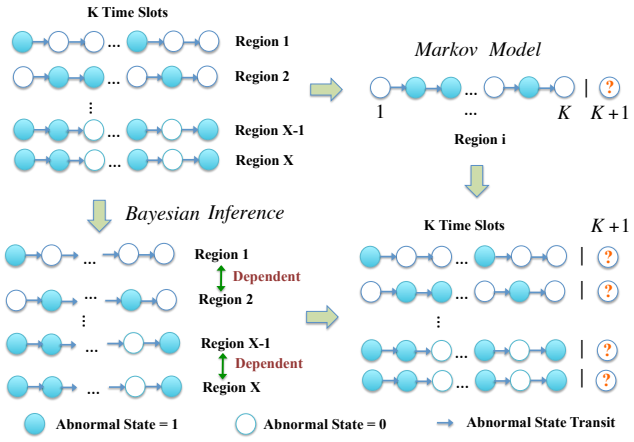

Figure 1: Crowdsourcing-based Urban Anomaly Prediction

3.1 Clustering Dependent Regions

We first cluster similar regions based on the number of anomalies and temporal vector features discussed in the previous section. In particular, we iteratively update hyperparameters of each cluster distribution (i.e., γ and η) and the cluster assignment (i.e., e) in the Bayesian inference model until convergence.

3.1.1 Updating Cluster Hyperparameters

To update cluster hyperparameters, we maximize the likelihood function $\Pr(Y, t_\Delta|\gamma, \eta, e)$ with respect to γ and η respectively. Note that adjusting γ_c only affects the likelihood with respect to Y_i of regions in cth cluster. Therefore, we are equivalently maximizing $\prod_{i,e_i=c} \Pr(Y_i|\gamma_c, e)$ and compute the maximum likelihood update for γ_c given the anomalies distribution Y_i of regions in cth cluster. Similarly, we are equivalently maximizing $\prod_{i,e_i=c} \Pr(t_{\Delta_i}|\eta_c, e)$ and compute the maximum likelihood update for η_c given the temporal distribution t_{Δ_i} of regions in cth cluster. Formally, the step of updating cluster hyperparameters can be written as follows:

$$\gamma_c^* = \arg\max_{\gamma_c} \Pr(Y|\gamma_c, e), \quad \eta_c^* = \arg\max_{\eta_c} \Pr(t_\Delta|\eta_c, e) \quad (2)$$

Finding the maximum likelihood updates for γ_c can be derived from the two-step process from a Dirichlet-multinomial distribution [8].

3.1.2 Updating Cluster Assignments

Based on the updates of cluster hyperparamters, we adjust the cluster assignment e_i by maximizing the likelihood function. Note that varying e_i only affects the likelihood with respect to region i. Thus, maximizing the likelihood

function $\Pr(Y, t_\Delta | e_i = c)$ is equivalent as:

$$e_i^* = \arg\max_c \Pr(Y_i | e_i = c) \Pr(t\_\Delta_i | e_i = c) \quad (3)$$

To compute $\Pr(Y_i | e_i = c)$, note that it is the probability of drawing Y_i from a Dirichlet-Multinomial distribution with know parameters γ_c and η_c. [8] provides the solutions for cluster assignments.

3.2 Markov Trajectory Estimation

We consider a the abnormal state transition of each region follows a Markov model. For simplicity, we assume the state in the predicted slot ($AS_{i,k+1}$) only depends on its previous state ($AS_{i,k}$) in this paper. We noted that we could also consider multiple previous states to predict the current state.

In a Markov model with binary variables (i.e., the value of $AS_{i,k}$ are binary), two transition probabilities are enough to describe the system dynamics: (i) $P_{0,1}$: the probability that $AS_{i,k}$ changes its state from 0 to 1 and (ii) $P_{1,0}$: the probability that $AS_{i,k}$ changes its state from 1 to 0. The probability $P_{0,0}$ that $AS_{i,k}$ remains in the 0 state in the next time-slot can be easily computed as $P_{0,0} = 1 - P_{0,1}$. Similarity, $P_{1,1} = 1 - P_{1,0}$.

Given a state trajectory, and the probability of its initial 0 state P_0^0 or 1 state P_1^0, we can compute its probability. For example, if the state trajectory is $1, 0, 1$, the joint probability of the state sequence is $P_1^0 \cdot P_{1,0} \cdot P_{0,1}$, where $P_{1,0}$ and $P_{0,1}$ are the transition probabilities. Then we use the transitional probabilities we learned from the last K time-slots to predict the state of time slot $K + 1$.

3.3 Optimal Anomaly Prediction

We define regions within the same cluster as dependent regions. For each region, we use Ne_i to represent the set of its dependent regions. Using the Markov model discussed above, we can estimate the next state (i.e., $K + 1$th time slot) of each region (i.e., $AK_{i,(K+1)}$). Based on the outputs of the above two components, we can estimate an optimized state value of each region by leveraging its dependency with other regions and its individual abnormal state trajectory. In particular, we define the objective function of our problem as follows:

$$f = \sum_{i=1}^{X} \sum_{i' \in Ne_i} |AK_{i,(K+1)}{}^* - AK_{i',(K+1)}|, \; AK_{i',(K+1)} \in [0, 1] \quad (4)$$

where i' is the index for the dependent regions of region i and $AK_{i,(K+1)}{}^*$ is the optimized inference of $(K + 1)$th state of region i. $AK_{i',(K+1)}$ is the state inference from Markov model. Here, the goal is to find the $AK_{i,(K+1)}{}^*$ for every region that minimizes the defined objective function. This optimization problem can be solved in linear time using weighted median algorithm.

4. EVALUATION

In this section, we conduct experiments to evaluate the performance of the *Crowdsourcing-based Urban Anomaly Prediction Scheme (CUAPS)* scheme on four real world datasets we collected from 311 service in New York City (NYC). We demonstrate the effectiveness of our proposed scheme on this dataset and compare the performance of our scheme to the state-of-the-art baselines. In the rest of this section: (i) we present the experiment settings and evaluation metrics we used in our experiments. (ii) We introduce the state-of-the-art baselines and present the evaluation results that demonstrate the *CUPAS* scheme can predict different categories of anomalies in a city more accurately than the compared baselines.

4.1 Experiment Setups and Evaluation Metrics

4.1.1 Dataset Statistics

311 is NYC's non-emergency service platform. This platform allows people to complain things happened around them by texting, phone call or mobile app. Each complaint record is formatted as: (complaint category, latitude, longitude, timestamp). The complaint categories in our four collected datasets are *Noise*, *Blocked Driveway*, *Illegal Conversion Of Residential Building (ICRB)* and *Illegal Parking* respectively. The time duration of the collected 311 complaints from Jan 2014 to Mar 2015. In NYC, we use road segments with a level from L_1 to L_5 as major roads to partition the entire city, which resulting in 862 regions [12]. Then each 311 complaint can be mapped to one of these regions. We note that the data is very sparse in an individual region, which makes the anomaly prediction problem a very challenging task. The statistics of these datasets are summarized in Table 1.

Table 1: Data Traces Statistics

Complain Category	Noise	Blocked Driveway	ICRB	Illegal Parking
Number of Instances	151174	92335	27724	69100

4.1.2 Evaluation Metric

In our evaluation, we use the following metrics to evaluate the estimation performance of the *CUAPS* scheme: *Precision*, *Recall*, *F1-measure* and *Accuracy*. Their definitions are given in Table 2.

Table 2: Metric Definitions

Metric	Definition
$Precison$	$\frac{TP}{TP+FP}$
$Recall$	$\frac{TP}{TP+FN}$
$F1-measure$	$\frac{2 \times Precison \times Recall}{Precison+Recall}$
$Accuracy$	$\frac{TP+TN}{TP+TN+FP+FN}$

4.2 Evaluation of Our Scheme

In this subsection, we evaluate the performance of the proposed *CUAPS* scheme and compare it to the state-of-the-art techniques as follows:

- *Bayesian Inference with Ground Truth (BIGT)*: it uses Bayesian model to cluster dependent regions and applies ground truth information of dependent regions to predict anomaly in a given region.
- *Markov Model (MM)*: it uses the Markov model to predict current state of the system from the transitional probabilities learned from the state trajectory of a region.

Table 3: Prediction Results on NY 311 Service Datasets

	Noise				Blocked Driveway			
Algorithm	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
CUAPS	**0.712**	**0.760**	**0.730**	**0.745**	**0.651**	**0.719**	**0.640**	**0.677**
BIGT	0.594	0.630	0.711	0.668	0.571	0.629	0.609	0.619
MM	0.645	0.738	0.596	0.659	0.589	0.645	0.625	0.634
GP	0.505	0.582	0.500	0.537	0.535	0.600	0.562	0.580
Random	0.461	0.536	0.467	0.500	0.446	0.517	0.468	0.491
	ICRB				Illegal Parking			
Algorithm	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
CUAPS	**0.655**	**0.667**	**0.823**	**0.736**	**0.627**	**0.644**	**0.906**	**0.753**
BIGT	0.586	0.666	0.588	0.625	0.549	0.595	0.875	0.708
MM	0.517	0.800	0.235	0.363	0.622	0.750	0.617	0.677
GP	0.482	0.550	0.647	0.594	0.603	0.782	0.529	0.631
Random	0.448	0.529	0.529	0.529	0.431	0.560	0.437	0.491

- *Gaussian Processing (GP)*: it is a nonparametric approach to predict the abnormal state of each region by exploring temporal features [1].

- *Random*: it randomly guesses the next abnormal state of each region.

4.2.1 Evaluation Results

In our evaluation, we evaluated the above schemes using the data from Jan 2014 to Sep 2014 to training data to predict the abnormal state of next day. To accommodate some baselines that need a small fraction of ground truth labels as inputs, we set the percent of ground to be used as 20%. The evaluation results of Noise, Blocked Driveway, Illegal Conversion Of Residential Building (ICRB) and Illegal Parking in Table 3. We observe that *CUAPS* outperforms the compared baselines in most of the evaluation metrics: it predicts correctly the most number of next abnormal states while keeping the falsely reported one the least.

5. CONCLUSION

In this paper, we developed a *Crowdsourcing-based Urban Anomaly Prediction Scheme (CUAPS)* to accurately predict urban anomalies for smart city applications. The CUAPS allows us to derive an *optimal* solution to accurately predict different categories of anomaly at different regions in a city. It explores both the dependency of anomaly occurrence between different regions and the historical anomaly distribution of an individual region. We evaluate our new scheme on four real-world datasets collected from 311 service in the city of New York. The results showed that our scheme outperforms state-of-the-art baselines in terms of prediction accuracy.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. CBET-1637251, CNS-1566465 and IIS-1447795 and Army Research Office under Grant W911NF-16-1-0388. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

6. REFERENCES

[1] P. Esling and C. Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, page 12, 2012.

[2] A. Ferraz Costa, Y. Yamaguchi, A. Juci Machado Traina, C. Traina Jr, and C. Faloutsos. Rsc: Mining and modeling temporal activity in social media. In *KDD*, pages 269–278. ACM, 2015.

[3] H.-P. Hsieh, S.-D. Lin, and Y. Zheng. Inferring air quality for station location recommendation based on urban big data. In *KDD*, pages 437–446. ACM, 2015.

[4] C. Huang and D. Wang. Exploiting spatial-temporal-social constraints for localness inference using online social media. In *ASONAM*. ACM/IEEE, 2016.

[5] C. Huang and D. Wang. Topic-aware social sensing with arbitrary source dependency graphs. In *IPSN*. ACM/IEEE, 2016.

[6] C. Huang and D. Wang. Unsupervised interesting places discovery in location-based social sensing. In *DCOSS*, pages 67–74. IEEE, 2016.

[7] J. Y. Lee, U. Kang, D. Koutra, and C. Faloutsos. Fast anomaly detection despite the duplicates. In *WWW*, pages 195–196. ACM, 2013.

[8] T. Minka. Estimating a dirichlet distribution, 2000.

[9] D. Wang, T. Abdelzaher, and L. Kaplan. Surrogate mobile sensing. *IEEE Communications Magazine*, 52(8):36–41, 2014.

[10] D. Wang and C. Huang. Confidence-aware truth estimation in social sensing applications. In *Sensing, Communication, and Networking (SECON), 2015 12th Annual IEEE International Conference on*, pages 336–344. IEEE, 2015.

[11] J. Wang, Y. Zhao, and D. Wang. A novel fast anti-collision algorithm for rfid systems. In *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, pages 2044–2047. IEEE, 2007.

[12] Y. Zheng, H. Zhang, and Y. Yu. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. *SIGSPATIAL*, 2015.