Towards Optimized Online Task Allocation in Cost-Sensitive Crowdsensing Applications

Yang Zhang, Daniel (Yue) Zhang, Qi Li, Dong Wang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
yzhang42@nd.edu, yzhang40@nd.edu,qli8@nd.edu,dwang5@nd.edu

***Abstract*—In crowdsensing applications, participants (crowd sensors) work collectively to report their measurements about the physical world. This paper focuses on the *optimized online task allocation* problem in cost-sensitive crowdsensing applications where the goal is to dynamically allocate the sensing tasks to participants to meet the requirement of the applications while minimizing the sensing costs. Recent progress has been made to tackle the task allocation problem in crowdsensing. However, two important challenges have not been well addressed: i) "physical dynamics": the values of the measured variables in crowdsensing often change significantly over time and space. It is essential for the task allocation schemes to adapt to such changes efficiently to optimize the task allocation process; ii) "crowd irregularity": the number of participants in crowdsensing is often smaller than the number of desirable sensing locations and not all crowd sensors contribute data all the time (e.g., due to incentive or budget constraints). To address the above challenges, this paper develops an Online Optimized Task Allocation (OO-TA) scheme inspired by techniques from information theory and online learning. We evaluate the OO-TA scheme using a dataset collected from a real-world crowdsensing application. The evaluation results show that OO-TA scheme significantly outperforms the state-of-the-art baselines in terms of both effectiveness and efficiency.**

***Keywords*-Crowdsensing, Physical Dynamics, Crowd Irregularity, Online Learning**

I. Introduction

In this paper, we develop a new analytical framework to address the online optimized task allocation problem in crowdsensing applications. In crowdsensing, participants (crowd sensors) work collectively to report their measurements about the physical world [1], [2], [3]. Examples of such applications include monitoring the air quality of a city using mobile phones [4], reporting malfunctioning urban infrastructures using geotagging [5], and obtaining real-time situation awareness in disaster response using inputs from common citizens [6].

In crowdsensing applications, the budget of sensing resource is often limited (e.g., a limited number of crowd participants or incentives). There exists a fundamental tradeoff between the sensing costs and the measurement errors: a higher sensing cost often allows for more crowd sensors to be involved in the application, which usually leads to a smaller overall measurement error. However, the high sensing cost may not always be affordable or necessary for a crowdsensing application (e.g., due to budget constraints or resource availability) [7]. Therefore, it is important to minimize the sensing cost while meeting the performance requirements (e.g., error bounds) of the application by choosing the *right set of tasks* for the crowd sensors. We refer to this problem as *optimized task allocation* in cost-sensitive crowdsensing applications.

A straightforward way to solve the above problem is to allocate crowd sensors to each subarea (which is referred to as a *cell* in the rest of the paper) and collect the sensing measurements. A major issue of this solution is that it encounters a high sensing cost of allocating tasks to crowd sensors in all cells [8], [4]. An alternative approach is to only select a few cells for sensing and infer the data for the rest of the cells from the sensed ones. This method is feasible since many physical sensing entities (e.g., noise, air quality, traffic) have strong spatial and temporal correlations that can be used for the inference. A rich set of solutions have been developed to address the optimized task allocation problem in crowdsensing [8], [4], [9], [10], [11]. They mainly focus on finding the best task allocation strategy subject to specific goals such as maximizing the sensing coverage [9], minimizing the sensing costs [8], or reducing the redundancy in the sensing data [4]. However, two important challenges have not been fully addressed by current solutions: *physical dynamics* and *crowd irregularity*, which are elaborated below.

Physical Dynamics. In crowdsensing, the readings of the measured variables often have large physical dynamics [4]. For example, Figure 1 shows the dynamics of the air quality readings (i.e., daily maximum mixing height (HMIX)) from a real-world crowdsensing application [1]. We can observe that the readings change with significant variations on both spatial and temporal dimensions. It is essential for the task allocation scheme to respond quickly to such large physical dynamics and adjust allocation strategies to meet the performance requirement of applications. Several task allocation schemes have been developed to address similar problems [8], [12]. However, two important limitations exist: i) current solutions often *sequentially* allocate the sensing task to cells (i.e., select one cell at a time) until the performance requirement is met, making it difficult to capture the dynamics of measured variables [8]; ii) existing solutions mainly focus on maximizing

978-1-5386-6808-5/18/$31.00 ©2018 IEEE

[1]https://sites.google.com/a/r-inla.org/stbook/datasets

the sensing coverage in order to catch the physical dynamics, which may lead to unnecessarily high sensing costs [12].

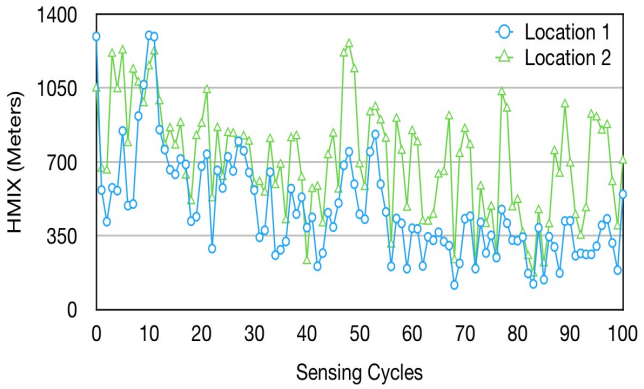


Fig. 1. Large Physical Dynamics in a Real-World Crowdsensing Application

Crowd Irregularity. The number of participants in crowdsensing is often smaller than the number of desirable sensing locations (e.g., due to incentive or budget constraints) and not all crowd sensors contribute data all the time (e.g., people may choose to participate the application based on their own schedule) [13]. The crowd irregularity issue makes the task allocation problem more challenging because it provides inadequate evidence to accurately identify the set of cells for optimized task allocations. A few task allocation methods have been developed to address this challenge [8], [4]. However, two important limitations exists: i) current task allocation methods are often tightly coupled with specific spatial-temporal inference algorithms and less generalizable to different crowdsensing applications [4], [8]; ii) many existing solutions do not fully incorporate cost-constraints into the optimization process of task assignments, leading to suboptimal task allocation results that do not minimize the overall sensing costs [4].

In this paper, we develop an Online Optimized Task Allocation (OO-TA) scheme to address above challenges. To address the *physical dynamics* challenge, our scheme efficiently captures the large dynamics of measured variables by making real-time decisions on how many and which cells should be chosen for task allocation in each sensing cycle. In particular, we develop an online learning algorithm to accurately estimate the optimal number of sensing cells that meets the performance requirement of the application with the minimum sensing cost. To address the *crowd irregularity* challenge, we develop a novel dynamic task priority ranking algorithm that estimates the priority of each cell for sensing task allocation by exploring the uncertainty of crowdsensing data from different perspectives. Our OO-TA scheme does *not* require any prior knowledge or performance guarantee from the inference algorithms and only expects the basic output (i.e., inferred sensing values from cells) and hence can be easily integrated with various inference schemes in crowdsensing applications. Finally, we evaluate the OO-TA scheme through a real-world crowdsensing application. The results show that OO-TA scheme significantly outperforms the state-of-the-art baselines in various application scenarios. In particular, our scheme reduces the inference error by 34.2% and reduces the sensing cost by 17.9% compared to the baselines.

II. Related Work

A. Crowdsensing

Crowdsensing is a powerful sensing paradigm where sensing measurements are collected from crowd sensors (either physical or human based) about the physical world [14], [15]. Examples of such applications include urban environment monitoring [4], target tracking [16], intelligent urban transportation [17], copyright infringing content detection [18], and real time disaster response [19]. There exist several important challenges in crowdsensing such as data fusion [20], networking and communication [21], data reliability [22], privacy protection [23], opinion characterization [24], system scalability [25], and incentives design [26]. Online optimized task allocation remains to be a significant but largely unsolved challenge in cost-sensitive crowdsensing applications [8]. The goal is to identify an optimal set of cells for sensing task allocation that meets the performance requirement of the application with a minimum sensing cost. This paper develops an Online Optimized Task Allocation (OO-TA) scheme to address this problem.

B. Task Allocation

Previous efforts have made good progress to address the task allocation problem in crowdsensing. For example, Wang *et al.* developed an active learning based task allocation approach to allocate the task with data quality guarantees [8]. Hsieh *et al.* developed a greedy based task allocation approach to deploy the sensors to the cells that are most likely to minimize the entropy generated by inference algorithms [4]. He *et al.* proposed a task allocation algorithm for location dependent crowdsensing tasks that maximizes the sensing coverage [9]. Ahmed *et al.* developed a spatial-temporal-aware allocation scheme that assigns tasks for a probabilistic sensing coverage [12]. In contrast, this paper develops a new task allocation scheme that addresses the *physical dynamics* and *crowd irregularity* challenges that have not been well addressed by the above solutions.

C. Online Learning

Our work is also related to online learning techniques which have been applied in social media, computer version, pattern detection and decision support [27], [28], [29], [30]. For example, Perozzi *et al.* developed a novel approach to learn latent social network representations via online learning [27]. Chechik *et al.* used an online approach to learn image similarity in large scale datasets [28]. Lattimore *et al.* presented algorithms of online learning method to predict optimal bus waiting time [29]. Kivinen *et al.* developed kernel based online learning algorithms for classification, regression and novelty detection [30]. To the best of our knowledge, OO-TA is among the firsts to address the optimized online task allocation problem in cost-sensitive crowdsensing applications using an online learning approach. In particular, we explicitly model the

optimized online task allocation problem as an online learning process and develop a novel dynamic control scheme to adjust the task allocation based on sensing measurements contributed by crowd sensors.

III. Problem Statement

In this section, we formulate the problem of online optimized task allocation in crowdsensing applications. We first define a few terms that will be used in the problem statement.

Definition 1: **Sensing Cell (**SCe**)**: We divide the target area for sensing task into disjoint cells where each cell represents a subarea of interest. We assume the sensing values are uniform in a sensing cell.

Definition 2: **Sensing Cycle (**SCy**)**: A sensing cycle is a period of time where crowd sensors perform one round of the sensing tasks.

Definition 3: **Ground Truth Sensing Value (**TS**)**: We define a TS matrix to represent the ground truth sensing value of all cells in all sensing cycles. In particular, $TS_{x,y}$ is the ground truth sensing value of cell x at cycle y.

Definition 4: **Collected Sensing Value (**CS**)**: We define a CS matrix to represent the collected sensing value of all cells in all sensing cycles. In particular, $CS_{x,y}$ is the collected sensing value of cell x at cycle y. $CS_{x,y} = TS_{x,y}$ if cell x is allocated for sensing task in cycle y. Otherwise, $CS_{x,y}$ is $null$.

Definition 5: **Reconstructed Sensing Value (**RS**)**: We define a RS matrix to represent the reconstructed sensing value of all cells in all sensing cycles. The reconstructed sensing values are often reconstructed by the inference algorithms from the collected sensing values. In particular, $RS_{x,y}$ is the reconstructed sensing value of cell x at cycle y.

Definition 6: **Sensing Error (**SE**)**: We define a SE matrix to represent the mean absolute error between the reconstructed sensing value and ground truth sensing value. In particular, $SE_{x,y} = |RS_{x,y} - TS_{x,y}|$. In addition, we define SE_y to be the average error of all cells at the cycle y and SE_{all} to be the average error for all cells over all cycles.

Definition 7: **Sensing Task Cost (**TC**)**: We define a TC vector to represent the sensing cost in all sensing cycles. In particular, TC_y represents the sensing cost of task allocation at cycle y. In this paper, we use the number of allocated tasks to represent the sensing cost.

Using the above definitions, let us consider a crowdsensing application with X sensing cells and Y sensing cycles. A set of N_y crowd sensors collectively report measurements of measured variables at cycle y. One sensor only reports measurements of one cell at a given sensing cycle. The number of sensors N_y is often significantly smaller than the number of sensing cell X due to the sensing budget and resource limitations. The task allocation scheme needs to identify a subset X_y cells from X cells for the sensing task allocation in each cycle to construct the collected sensing value matrix CS. Finally, the application infers the sensing values of the unsensed cells to construct the RS matrix and compute the sensing error in SE.

The goal of the online optimized task allocation is to make real-time task allocation decisions that ensure the required data quality of the application while minimizing the total sensing cost. We adopt a metric, (ϵ, p)-quality, to evaluate the sensing data quality in crowdsensing using a Bayesian inference approach [8]. The idea of (ϵ, p)-quality is to keep the sensing errors of at least p fraction of all sensing cycles to be less than a predefined error bound ϵ. For example, in a crowdsensing application to measure the temperature over 100 sensing cycles, ($0.5^\circ C$,0.8)-quality means at least 80 sensing cycles have the sensing errors less than $0.5^\circ C$. Based on the above definitions, our problem is formally defined as:

$$\begin{aligned} &\text{Select } X_y \text{ from } X \text{ to } min\sum_{y=1}^{Y_c} TC_y \\ &\text{while keeping } |\{y|SE_y < \epsilon\}| \geq Y_c \cdot p; \quad y \in [1, Y_c] \end{aligned} \tag{1}$$

where Y_c denotes the current sensing cycle. X_y is defined as the *optimal set* of cells selected for task allocation at cycle y that minimize the accumulative sensing costs while meeting the (ϵ, p)-quality requirement of the application.

IV. solution

In this section, we present the Online Optimized Task Allocation (OO-TA) scheme to address the problem formulated in the previous section. The OO-TA scheme estimates an optimal set of cells for sensing task allocation at each cycle in real-time and recursively updates the optimal set based on the sensing results from the previous cycles. The OO-TA scheme consists of two components: i) Dynamic Task Priority Ranking (DTPR) to determine the *priority of cells* for sensing task allocation given the incomplete spatial temporal sensing coverage; ii) Online Task Number Estimation (OTNE) to estimate the *size of the optimal set* and select the top ranked cells to be included in optimal set accordingly. The details of these components are presented as follows.

A. Dynamic Task Priority Ranking (DTPR)

In this subsection, we first present the Dynamic Task Priority Ranking component that computes the priority of a cell to be included in the optimal set for the sensing task allocation at each cycle. Based on the priority ranking list, the Online Task Number Estimation component can then select the top ranked cells in the optimal set.

To compute the priority of cells, we need to know which cell, whose sensing value if collected, would be the most helpful one to improve the overall sensing quality. However, this is difficult without knowing the true sensing value of cells in advance [8]. In crowdsensing, an alternative strategy is to allocate the sensing tasks to cells based on the uncertainty from the inference algorithms on that cell [4]. However, it is not a trivial task to accurately quantify the uncertainty of a cell given the incomplete spatial-temporal sensing coverage [4]. In our solution, we develop an estimation algorithm that estimates the priority of a cell based on two types of uncertainties that are directly related to the sensing error: *temporal uncertainty*

(TU) and *staleness uncertainty (SU)*, and dynamically update them in each sensing cycle.

Temporal Uncertainty (TU): We define the temporal uncertainty of a cell to be the variance of the inferred sensing values of the cell in all sensing cycles. A high variance of a cell indicates the inference algorithm may not be very confident about its inferred sensing values of that cell. Based on the temporal inference variance, we define the *temporal uncertainty* formally as follows:

$$TU_x^y = \frac{var(\{RS_{x,1}, RS_{x,2}, ..., RS_{x,y}\})}{\sum_{k=1}^{X} var(\{RS_{k,1}, RS_{k,2}, ..., RS_{k,y}\})} \quad (2)$$

where TU_x^y represents the temporal uncertainty for cell x in cycle y. $RS_{x,y}$ is the reconstructed sensing value of cell x at cycle y (Definition 5). $var(\{RS_{x,1}, RS_{x,2}, ..., RS_{x,y}\})$ is the function to calculate the temporal inference variance of cell x at current cycle y. X is the number of cells in the target area. Intuitively, the cell with higher temporal inference variance has the higher temporal uncertainty and vice versa.

Staleness Uncertainty (SU): We define the staleness uncertainty of a sensing cell as the idle time since the cell has been sensed last time. A long idle time in a cell indicates the sensing values in that cell are probably stale and the cell may lack sufficient fresh data for the inference algorithm to make an accurate inference in that cell. Specifically, we increase the staleness uncertainty of a cell when that cell is not selected for sensing in a cycle and set the staleness to an initial value when the cell is selected for sensing. Formally, we define the staleness uncertainty as follows:

$$SU_x^y = \begin{cases} 1, \text{ if the cell } x \text{ is selected in cycle } y \\ \gamma \cdot SU_x^{y-1}, \text{ otherwise} \end{cases} \quad (3)$$

where SU_x^y represents the staleness uncertainty for cell x at cycle y and it is set to 1 at the first cycle. γ is a predefined scaling factor (which is large than 1) that is used to increase the staleness uncertainty if a cell is not selected for task allocation. In particular, a larger γ value indicates a faster uncertainty accumulation speed for the non-selected cells and vice versa. The actual value of γ depends on the specific application.

We then combine the temporal uncertainty and staleness uncertainty into an *overall uncertainty (OU)* to compute the priority of cells as follows:

$$OU_x^y = \lambda_{TU}^y \times TU_x^y + \lambda_{SU}^y \times SU_x^y \quad (4)$$

where OU_x^y represents the overall uncertainty for cell x at cycle y. λ_{TU}^y and λ_{SU}^y is the weight for temporal uncertainty and staleness uncertainty at cycle y, respectively. The values of λ_{TU}^y and λ_{SU}^y are tuned for a given sensing application.

B. Online Task Number Estimation (OTNE)

In this subsection, we present the Online Task Number Estimation component of OO-TA scheme that estimates the size of the optimal set that satisfies the data quality requirement from the application with minimized sensing cost. In particular, we explicitly model the real-time task allocation as an online learning process where the OTNE component estimates the number of cells in the optimal set and dynamically updates this estimation based on the "feedback" received from the previous cycles. The feedback signal is driven by the difference between the minimum number of cells required to meet the data quality requirement of the application (we refer to such number as *optimal task allocation number*, k') and the actual number of sensed cells (i.e., the number of allocated tasks, k). Our goal is to minimize the difference between k' and k in each sensing cycle. In particular, we define a loss function ℓ as:

$$\ell = abs|k - k'| \quad (5)$$

where $abs||$ to be the absolute value between k' and k .

To minimize the loss function, we first sample the number of allocated tasks at each sensing cycle from a distribution $\mathcal{P}$ on the optimal task allocation number. In particular, we have:

$$p_y(x) = \begin{cases} (1-\eta) \cdot \frac{w_y(x)}{\sum_{i \in \{1,2,...,X\}} w_y(i)}, \text{ if } x \text{ not equals } X \\ (1-\eta) \cdot \frac{w_y(x)}{\sum_{i \in \{1,2,...,X\}} w_y(x)} + \eta, \text{ if } x \text{ equals } X \end{cases} \quad (6)$$

where $p_y(x)$ represents the probability of the optimal task allocation number k' to be x cells. η is a smoothing factor, which is usually set to be a small positive value. $w_y(x)$ represents the weight for the possible value x of being the optimal task allocation number at cycle y.

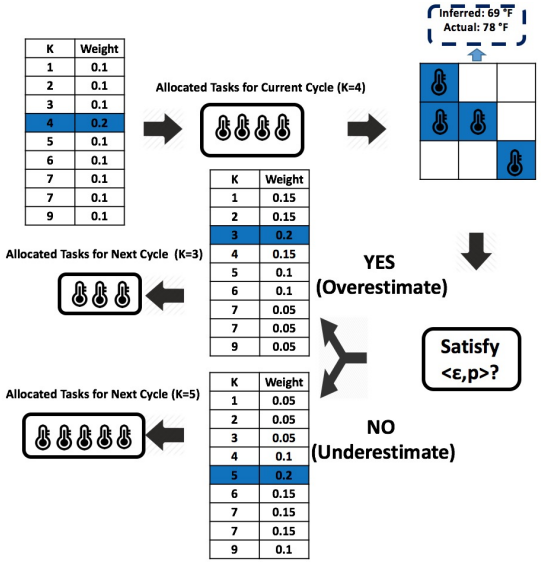
To update $w_y(x)$ in each sensing cycle y, we develop a dynamic updating scheme as follows.

$$w_y(x) = \Phi(w_{y-1}(x), k, k') \quad (7)$$

where Φ is a weight update function that dynamically updates the weight $w_y(x)$ that minimize the loss function defined in Equation 5. In this paper, we develop an enhanced Exp3-Dom algorithm [31] as the weight update function. The main idea of our scheme is to dynamically predict the weights of a set of *actions* using an exponentially weighted function based on the feedback from a loss function. An action with a higher weight is more likely to be selected. In particular, we define each possible number of cells for sensing task allocation (i.e., $k = 1, 2, ..., X$) as an action. The loss function ℓ we defined above is used as the feedback. Our scheme can then dynamically update the weight of each action by minimizing the loss function in real time (i.e., identifying the k that is closest to the optimal task allocation number k').

The online task number estimation scheme is guaranteed to find the optimal number of sensing tasks k' when the number of sensing cycle is big enough. Our scheme has been proved to achieve an optimal total loss bounded by logarithmic factors [31]. In particular, the total loss is bounded by $O(\sqrt{x})$ where x is the number sensing cycles. Therefore, the average loss ℓ in each sensing cycle is $O(1/\sqrt{x})$ (i.e., divide the total loss by the number of sensing cycles). We can observe that the average loss decreases quickly as the number of the sensing cycles increases. In another word, the number of allocated tasks estimated by our scheme (i.e., k) will converge to the optimal task allocation number (i.e., k') as we obtain data from more sensing cycles.

A simple illustrative example of OTNE scheme is shown in Figure 2. We have 9 cells in this example. At the current sensing cycle, the number of allocated sensing tasks is 4 (i.e., the k value with the highest weight). The data is collected from the 4 selected cells and the values of the remaining 5 cells are inferred from the collected sensing data. The data quality of current cycle is evaluated by using the (ϵ, p)-quality bound we introduced in Section III. If the quality bound is satisfied (i.e., we either overestimate or hit the value of optimal k), we increase the weights of the k values that are smaller than 4 for the next sensing cycle. For example, we allocate 3 sensing tasks (i.e., the k has the highest weight) in the next cycle. The intuition is that a k value that is smaller than the current selection might also satisfy the quality bound with a lower sensing cost considering the overestimation of k in the current cycle. If the quality bound is not satisfied, we will increase the weights of k values that are larger than the current selection (e.g., we allocate 5 tasks in next cycle) to increase the chance that the bound will be met in the next sensing cycle. The weights of different k values are adjusted by an exponential weights algorithm that dynamically adjusts the weights of k values to reach the optimal k [32].



K	Weight
1	0.1
2	0.1
3	0.1
4	0.2
5	0.1
6	0.1
7	0.1
7	0.1
9	0.1

K	Weight
1	0.15
2	0.15
3	0.2
4	0.15
5	0.1
6	0.1
7	0.05
7	0.05
9	0.05

K	Weight
1	0.05
2	0.05
3	0.05
4	0.1
5	0.2
6	0.15
7	0.15
7	0.15
9	0.1

Fig. 2. An Illustrative Example of Online Task Number Estimation (Sensing variable: temperature)

The online task number estimation scheme is shown in Algorithm 1. The input to the algorithm is the collected sensing values from cells in a sensing cycle, and the output is the estimated task allocation number k and the estimated set of cells X_y for task allocation in cycle y.

V. Evaluation

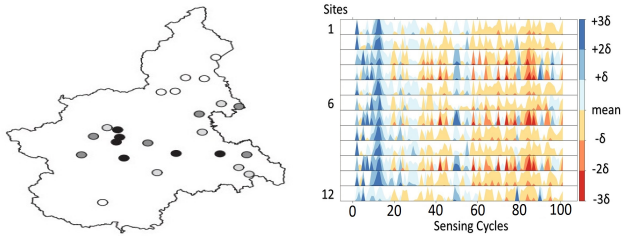
In this section, we evaluate the performance of the OO-TA scheme using the dataset collected from real world crowdsensing application: Piemonte Air pollution dataset. We compare the performance of the OO-TA scheme with the state-of-the-art task allocation baselines and the evaluation results show that the OO-TA significantly outperforms the compared baselines.

Algorithm 1 Online Task Number Estimation (OTNE)

for all $x \in \{1, 2, ..., X\}$ **do**
  $w_0(x) \leftarrow 1$
end for
$y \leftarrow 0$
while sensing cycles not end **do**
  $y \leftarrow y + 1$
  estimate the probability $p_y(X)$ for each x in distribution $\mathcal{P}$ using Equation 6
  sample the number of allocated task k from $\mathcal{P}$
  select top k ranked cells from PL_y as X_y
  for cells in X_y **do**
    collect sensing value CS_y
  end for
  $k' \leftarrow k$
  while $k' > 0$ **do**
    select first k' values from CS_y as CS'_y
    if CS'_y satisfies the $(\epsilon, p) - quality$ **then**
      $k' \leftarrow k' - 1$
    else
      the optimal k' is found
      break
    end if
  end while
  $k' \leftarrow k' + 1$
  if $k' > k$ **then**
    set $k' \leftarrow k$
  end if
  for $x \in (0, 1, 2, ..., X)$ **do**
    update $w_y(x)$ with k and k' using using Equation 7
  end for
end while

A. Datasets

In our evaluation, we use a real-world crowdsensing dataset published by Blangiardo *et al.* [33] [2]. This dataset contains air quality monitoring measurements collected from 24 locations in Piemonte region, Italy (see Figure 3(a)). There are 6 different daily air quality parameters in this dataset: PM10, daily maximum mixing height (HMIX), total precipitation, wind speed, temperature, and daily emission rates of primary aerosols. In our experiment, we choose the daily HMIX reading as the measured variable because HMIX is a critical environmental factor with a large spatial and temporal dynamics as shown in Figure 3(b)[3]. In this dataset, the sensing cycle is chosen to be one day.



(a) Spatial Distribution (b) Spatial-Temporal Dynamics

Fig. 3. Piemonte Air Quality Dataset

[2]https://sites.google.com/a/r-inla.org/stbook/datasets

[3]The dark colors in each horizontal line indicate large temporal dynamics and the various color patterns across different lines indicate large spatial variations.

B. Inference Algorithm

In the experiment, we couple four well-known inference algorithms with our OO-TA scheme to investigate the performance and robustness of the OO-TA scheme over different inference algorithms.

1) **Spatial k-Nearest Neighbour (Spatial kNN):** Spatial kNN infers the missing value of a cell by averaging the collected values from k nearest sensing cells [34].
2) **Inverse Distance Weighting (IDW):** IDW infers the missing value of a cell as the weighted mean of the collected values from the k nearest cells where the weights are inversely proportional to the distance between cells [35].
3) **Ridge Regression:** Ridge Regression, also know as Weight Decay, is a regression method that infers the missing value of a cell by reducing the multicollinearity in prediction variables [36].
4) **Support Vector Regression (SVR):** SVR is the regression version of the supervised learning method SVM that utilizes the collected data to build a prediction model to infer the missing value of a cell [37].

C. Baseline algorithms

While there exist a rich set of literature on addressing the task allocation problem in crowdsensing [8], [4], [12], [38], we choose several representative allocation schemes that are applicable to our online optimized task allocation problem as the baselines in the experiment.

1) **QBC**: QBC (Query by Committee Task Allocation) is an active learning based task allocation method that decides which cells to allocate tasks by using a set of inference algorithms to infer the sensing value for all cells. The cell with the largest variance among the inferred values of different algorithms will be selected first [38].
2) **Enhanced RND**: ERND (Enhanced Random Task Allocation) is a task allocation scheme used as a baseline in [8] that randomly selects the task number for sensing task allocation in each cycle. We enhance the RANDTA in our experiment by leveraging the results from the *Dynamic Task Priority Ranking* component to decide which cell will be selected first.
3) **GPR**: GPR (Greedy Priority Ranking Task Allocation) is a greedy based allocation algorithm that selects a fixed number of tasks in each cycle based on their priorities [4]. The priorities of cells are generated by the *Dynamic Task Priority Ranking* component.
4) **UNS**: UNS (Uniform Sampling Task Allocation) is a task allocation algorithm that decides the number o tasks from a uniform distribution and randomly selects the cells from the target area [39].

D. Evaluation Metrics

We use the following metrics to evaluate the task allocation performance of all compared schemes.

1) **Inference Error (IE):** We define the inference Error as the mean absolute error (MAE) between the reconstructed (inferred) sensing values and collected sensing values of all cells. Specifically, we define:

$$IE = \frac{\sum_{y=1}^{Y} SE_y}{Y} \tag{8}$$

where SE_y is the average error of all cells at cycle y as we defined in *Definition 6* in Section III and Y is the total number of cycles.

2) **Data Quality (DQ):** We define the data quality as the fraction of sensing cycles that have the mean sensing error to be smaller than the predefined error bound. This metric is used to evaluate whether our OO-TA scheme could always meet the data quality requirement from the application. Specifically, we define:

$$DQ = \frac{|\{y|SE_y <= \epsilon, 1 < y < Y\}|}{Y} \tag{9}$$

where the ϵ is the predefined error bound.

3) **Sensing Task Cost (TC_{all}):** We use the number of allocated tasks to represent the sensing costs as we discussed in Section III. Specifically, we define:

$$TC_{all} = \frac{\sum_{y=1}^{Y} TC_y}{Y} \tag{10}$$

to be average sensing cost in a cycle where TC_y is the sensing cost at cycle y.

4) **Optimal Task Allocation Error (AE):** We define the optimal task allocation error to be the mean absolute error (MAE) between the number of allocated sensing tasks and the optimal number of sensing tasks (i.e., the minimum number of tasks that meet the data quality requirement). Specifically, we define:

$$AE = \frac{\sum_{y=1}^{Y} |TC_y - \bar{TC}_y|}{Y} \tag{11}$$

where $\bar{TC}_y$ is the optimal number of sensing tasks.

E. Evaluation Results

In this subsection, we present the results of all compared schemes through extensive experiments on the real world dataset. In particular, we evaluate the performance of all schemes using the evaluation metrics discussed above. For the Piemonte Air quality dataset, we have 24 sensing cells and 100 sensing cycles. The data quality metric is set as (180 meters, 0.8)-quality. In the experiment, we choose the number of allocated tasks for each baseline in a sensing cycle that optimizes the performance of the corresponding baseline for a fair comparison.

1) Inference Error: We first evaluate the performance of the task allocation schemes in terms of their Inference Errors (IE). Figure 4 shows the inference errors of all task allocation schemes when they are coupled with different inference algorithms on Piemonte Air Quality dataset. We observe that the OO-TA scheme outperforms all baselines with the lowest inference errors. This is because the OO-TA scheme dynamically estimates the optimal set of cells that reduces

the inference error to the maximum extent within a certain sensing cost budget. We also observe the performance gain of the OO-TA scheme is consistent when it is coupled with different inference algorithms.

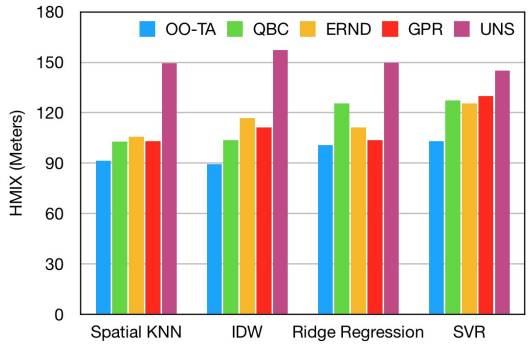


Fig. 4. Inference Error Results

2) Data Quality: We further evaluate whether our OO-TA scheme is able to meet the predefined (ϵ, p)-quality bound. Figure 5 shows the Data Quality (DQ) of the OO-TA scheme with different inference algorithms on Piemonte Air Quality dataset. We observe that the OO-TA scheme always meets the predefined (ϵ, p)-quality bound: at least 0.8 fraction of the sensing cycles have their sensing errors to be less than the predefined ϵ. This is because our OO-TA scheme explicitly considers the data quality of the application in its task allocation model and assigns tasks to collect the necessary data to meet the desired (ϵ, p)-quality.

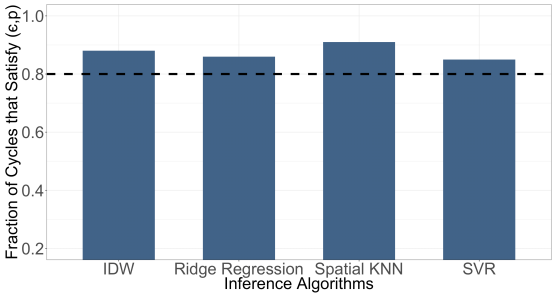


Fig. 5. Data Quality Results

3) Sensing Task Cost: We continue to evaluate the performance of the task allocation schemes on the sensing task cost metric TC_{all}. Figure 6 show the sensing task costs (i.e., the number of allocated tasks) of all compared schemes on Piemonte Air Quality dataset. We can observe that the OO-TA scheme outperforms all baselines over different inference algorithms with the lowest average sensing cost. This is because the OO-TA scheme consists of an online learning algorithm that dynamically estimates the *minimum* required number of sensing tasks in each sensing cycle to meet the data quality requirement of the application.

4) Optimal Task Allocation Error: Finally, we evaluate whether our OO-TA scheme is able to accurately estimate the optimal number of cells for the task allocation. In particular, we study the performance of the task allocation schemes using

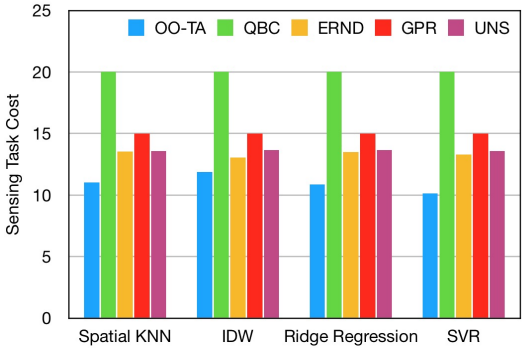


Fig. 6. Sensing Task Cost (Unit: Cells)

the Optimal Task Allocation Error metric AE. Figure 7 show the optimal task allocation error of different task allocation schemes on Piemonte Air Quality dataset. Our OO-TA scheme outperforms all baselines over different inference algorithms with the lowest errors. The results demonstrate that our OO-TA scheme can accurately estimate the optimal number of cells for task allocation, which directly contributes to the high inference accuracy and low sensing costs as shown earlier.

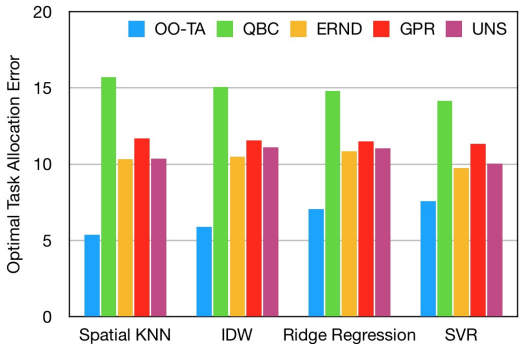


Fig. 7. Optimal Task Allocation Error (Unit: Cells)

This concludes the evaluation section. The OO-TA scheme is shown to outperform the compared baselines over different evaluation metrics when it is coupled with different inference algorithms. Such performance gains are mainly achieved by the DTPR and OTNE components of the scheme (discussed in Section IV) that collectively decide the optimal set of tasks allocated for sensing in an online fashion.

VI. Conclusion

This paper develops a novel OO-TA scheme to solve the online optimized task allocation problem in cost-sensitive crowdsensing applications. The OO-TA scheme addresses two fundamental challenges: *physical dynamics* and *crowd irregularity*. In particular, we develop a dynamic task priority ranking algorithm to decide the priority of each cell for the task allocation and an online learning algorithm to estimate the optimal number of sensing cells that meets the data quality requirement of the application. The evaluation results on a real-world dataset demonstrate that the OO-TA scheme achieves significant performance gains compared to the baselines.

ACKNOWLEDGEMENT

This research is supported in part by the National Science Foundation under Grant No. CNS-1831669, CBET-1637251, CNS-1566465, IIS-1447795, Army Research Office under Grant W911NF-17-1-0409, Google 2017 Faculty Research Award. 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.

REFERENCES

[1] J. A. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," *Center for Embedded Network Sensing*, 2006.

[2] D. Y. Zhang, C. Zheng, D. Wang, D. Thain, X. Mu, G. Madey, and C. Huang, "Towards scalable and dynamic social sensing using a distributed computing framework," in *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 2017, pp. 966–976.

[3] D. Wang, T. Abdelzaher, and L. Kaplan, *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann, 2015.

[4] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "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, 2015, pp. 437–446.

[5] J. Zhang and D. Wang, "Duplicate report detection in urban crowdsensing applications for smart city," in *Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on*. IEEE, 2015, pp. 101–107.

[6] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Information Processing in Sensor Networks (IPSN), 2012 ACM/IEEE 11th International Conference on*. IEEE, 2012, pp. 233–244.

[7] L. G. Jaimes, I. Vergara-Laurens, and M. A. Labrador, "A location-based incentive mechanism for participatory sensing systems with budget constraints," in *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on*. IEEE, 2012, pp. 103–108.

[8] L. Wang, D. Zhang, A. Pathak, C. Chen, H. Xiong, D. Yang, and Y. Wang, "Ccs-ta: Quality-guaranteed online task allocation in compressive crowdsensing," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 683–694.

[9] S. He, D.-H. Shin, J. Zhang, and J. Chen, "Toward optimal allocation of location dependent tasks in crowdsensing," in *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2014, pp. 745–753.

[10] H. Xiong, D. Zhang, L. Wang, and H. Chaouchi, "Emc 3: Energy-efficient data transfer in mobile crowdsensing under full coverage constraint," *IEEE Transactions on Mobile Computing*, vol. 14, no. 7, pp. 1355–1368, 2015.

[11] Y. Zhang, N. Vance, D. Zhang, and D. Wang, "Optimizing online task allocation for multi-attribute social sensing," in *The 27th International Conference on Computer Communications and Networks (ICCCN 2018)*. IEEE, 2018.

[12] A. Ahmed, K. Yasumoto, Y. Yamauchi, and M. Ito, "Distance and time based node selection for probabilistic coverage in people-centric sensing," in *Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2011 8th Annual IEEE Communications Society Conference on*. IEEE, 2011, pp. 134–142.

[13] D. Zhang, Y. Ma, Y. Zhang, S. Lin, X. Hu, and D. Wang, "A real-time and non-cooperative task allocation framework for social sensing applications in edge computing systems," in *Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2018.

[14] A. Ghasemi and E. S. Sousa, "Opportunistic spectrum access in fading channels through collaborative sensing." 2007.

[15] D. Y. Zhang, R. Han, D. Wang, and C. Huang, "On robust truth discovery in sparse social media sensing," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1076–1081.

[16] E. Xu, Z. Ding, and S. Dasgupta, "Target tracking and mobile sensor navigation in wireless sensor networks," *IEEE Transactions on mobile computing*, vol. 12, no. 1, pp. 177–186, 2013.

[17] S. Ilarri, O. Wolfson, and T. Delot, "Collaborative sensing for urban transportation." *IEEE Data Eng. Bull.*, vol. 37, no. 4, pp. 3–14, 2014.

[18] D. Y. Zhang, Q. Li, H. Tong, J. Badilla, Y. Zhang, and D. Wang, "Crowdsourcing-based copyright infringement detection in live video streams," in *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2018*, 2018.

[19] M. T. Al Amin, T. Abdelzaher, D. Wang, and B. Szymanski, "Crowd-sensing with polarized sources," in *Distributed Computing in Sensor Systems (DCOSS), 2014 IEEE International Conference on*. IEEE, 2014, pp. 67–74.

[20] P. Giridhar, M. T. Amin, T. Abdelzaher, D. Wang, L. Kaplan, J. George, and R. Ganti, "Clarisense+: An enhanced traffic anomaly explanation service using social network feeds," *Pervasive and Mobile Computing*, vol. 33, pp. 140–155, 2016.

[21] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed diffusion: A scalable and robust communication paradigm for sensor networks," in *Proceedings of the 6th annual international conference on Mobile computing and networking*. ACM, 2000, pp. 56–67.

[22] D. Wang, L. Kaplan, and T. F. Abdelzaher, "Maximum likelihood analysis of conflicting observations in social sensing," *ACM Transactions on Sensor Networks (ToSN)*, vol. 10, no. 2, p. 30, 2014.

[23] C. Cornelius, A. Kapadia, D. Kotz, D. Peebles, M. Shin, and N. Triandopoulos, "Anonysense: privacy-aware people-centric sensing," in *Proceedings of the 6th international conference on Mobile systems, applications, and services*. ACM, 2008, pp. 211–224.

[24] Y. Zhang, N. Vance, D. Zhang, and D. Wang, "On opinion characterization in social sensing: A multi-view subspace learning approach," in *Distributed Computing in Sensor Systems (DCOSS), 2018 International Conference on*. IEEE, 2018.

[25] D. Zhang, D. Wang, N. Vance, Y. Zhang, and S. Mike, "On scalable and robust truth discovery in big data social media sensing applications," *IEEE Transactions on Big Data*, 2018.

[26] D. Zhang, Y. Ma, C. Zheng, X. Hu, and D. Wang, "Cooperative-competitive task allocation in edge computing for delay-sensitive social sensing," in *Proceedings of the Third ACM/IEEE Symposium on Edge Computing (SEC)*, 2018.

[27] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.

[28] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *Journal of Machine Learning Research*, vol. 11, no. Mar, pp. 1109–1135, 2010.

[29] T. Lattimore, A. György, and C. Szepesvári, "On learning the optimal waiting time," in *International Conference on Algorithmic Learning Theory*. Springer, 2014, pp. 200–214.

[30] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE transactions on signal processing*, vol. 52, no. 8, pp. 2165–2176, 2004.

[31] N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir, "Nonstochastic multi-armed bandits with graph-structured feedback," *arXiv preprint arXiv:1409.8428*, 2014.

[32] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and computation*, vol. 108, no. 2, pp. 212–261, 1994.

[33] M. Blangiardo and M. Cameletti, *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

[34] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[35] R. Franke, "Scattered data interpolation: tests of some methods," *Mathematics of computation*, vol. 38, no. 157, pp. 181–200, 1982.

[36] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," 1998.

[37] K. Chen and J. Yu, "Short-term wind speed prediction using an unscented kalman filter based state-space support vector regression approach," *Applied Energy*, vol. 113, pp. 690–705, 2014.

[38] S. Chakraborty, J. Zhou, V. Balasubramanian, S. Panchanathan, I. Davidson, and J. Ye, "Active matrix completion," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 81–90.

[39] C.-J. Ho and J. W. Vaughan, "Online task assignment in crowdsourcing markets." in *AAAI*, vol. 12, 2012, pp. 45–51.