Data-Driven Similarity-based Worker Recruitment Towards Multi-task Data Inference for Sparse Mobile Crowdsensing

En Wang[†], Zijie Tian[†], Yongjian Yang[†], Wenbin Liu[†], Baoju Li[†], Nan Jiang[‡] and Jie Wu[§]
[†]Department of Computer Science and Technology, Jilin University, China
[‡]East China Jiaotong University, China
[§]Temple University, USA
Email: {wangen,yyj,liuwenbin}@jlu.edu.cn, {tianzj21,libj21}@mails.jlu.edu.cn, jiangnan@ecjtu.edu.cn, jiewu@temple.edu

***Abstract*—Sparse Mobile Crowdsensing is an emerging paradigm for data collection with budgets and workers' limitations, which recruits workers to sense a part of spatio-temporal data and infer what is unsensed. In order to achieve high inferring accuracy in all spatio-temporal areas, we need to measure the importance level of each area and sense some important ones. Existing works usually use the average distance or the difficulty level inferred by historical data to measure the area's importance. However, we argue that neither distance nor difficulty level is suitable for measuring the importance. A better approach is inspired by the data itself, i.e., data similarity among different areas. Furthermore, there usually exist multiple data types in sparse mobile crowdsensing, which leads to a more complex inference from two-dimensional data (spatial and temporal) to three-dimensional data (spatial, temporal, and data type). In this paper, we study worker recruitment in a multi-task scenario, which aims to recruit workers to sense important data for a three-dimensional inference. Specifically, we first present the SWDTW method to calculate data similarity, which is used to evaluate data importance. Based on this, we further propose an evaluation model for three-dimensional data similarity and measure the importance of each area. Finally, inspired by generalized greedy and simulated annealing, we propose a worker recruitment method named WRGSA, the target of which is selecting workers to sense important areas to enhance the inference accuracy. Extensive experiments have been conducted over three real-world datasets with multiple data types, which can verify the effectiveness of our proposed methods.**

***Index Terms*—Sparse MCS, Similarity, Multiple data type, Worker recruitment**

I. Introduction

Mobile Crowdsensing (MCS) [1] is an efficient way of carrying out data collection, which can recruit users who are equipped with smart devices to collect various data. It has been widely used in many applications, including environment monitoring [2], traffic management [3], urban monitoring [4], etc. The common concern of the above applications focuses on achieving high accuracy data with a limited sensing budget, which raises Sparse MCS [5], aiming at inferring unsensed areas using a little sensed data.

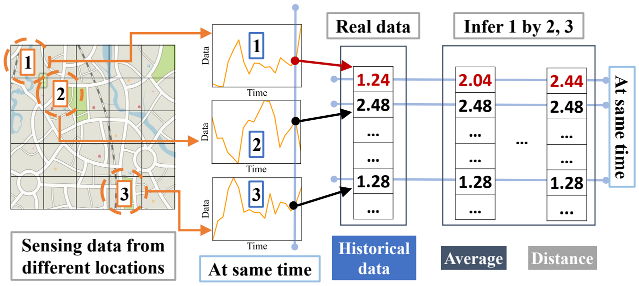


Fig. 1: Problem description of data inferring methods

In practice, if we manage to sense some more important areas, then we can infer the rest of the areas with high accuracy. As a consequence, how to evaluate the importance level of each area becomes an essential issue in Sparse MCS. To evaluate such importance, some researchers [6] use the average distance between different spatio-temporal as the information entropy. Some other works use the difficulty level, such as the numerical difference between sensing cycles [7], and the variance of multiple inference algorithms [8] to measure the inference importance. However, we argue that all of these works are not suitable to measure the importance in some cases. For example, as shown in Fig.1, there are 16 areas of which three areas' historical data are listed. The distance between areas 1 and 2 is close while their time-varying data is totally different. In contrast, the data of areas 1 and 3 is very similar even if these two areas are far from each other. So when inferring area 1, the data of area 3 is more valuable. Inspired by this phenomenon, a better way may be using the similarity of time-varying data [9] instead of the distance of two areas to measure the inference importance. When evaluating the similarity, the imprecisely inferred data may mislead the measurement of similarity, so we must take this into account when we calculate the similarity. Therefore, how to calculate such similarity according to the historical data is the first challenge in this paper.

Moreover, in mobile crowdsensing, there may exist a myriad of types of data to be collected (temperature, humidity, light, etc), and such types of data can also be inferred from the other

Corresponding author: Wenbin Liu.

979-8-3503-9973-8/23/$31.00 © 2023 IEEE

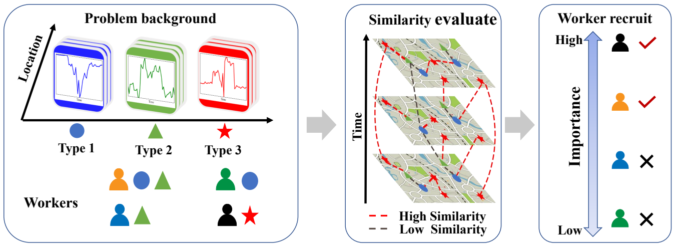


Fig. 2: Worker recruitment with similarity.

ones. However, many existing works of data importance evaluation are designed either for the single data-type task scenario, or for the multiple type scenario but only consider that tasks are independent from each other, which ignores the inference among different types of data. Such phenomena inspire us to consider the importance in multiple data-type scenarios, which makes importance evaluation more difficult because the data inference changes from the two-dimensional data (spatial and temporal) to three-dimensional data (spatial, temporal and data type). Hence, how to measure such similarity and use it to evaluate data importance in such three-dimensional data is the second challenge which urgently needs to be solved.

Finally, even if we could decide the important areas for data inference, we still need to select some suitable workers to sense the corresponding data, which requires us evaluating each worker's importance. In practice, a worker may carry different sensors which can collect different types of data. For a specific crowdsensing task, limited budget, and user's trace coverage and the equipped sensor types make the user selection become more difficult. Based on the above consideration, how to select a group of suitable workers is the third challenge.

To solve the three challenges above, we focus on the data similarity under the multiple data-types scenario, and propose a powerful method to recruit suitable workers to sense important data for a three-dimensional inference. The definition of our proposed problem is shown in Fig.2. The left of Fig.2 presents the historical distribution of three types of tasks and four workers which are equipped with specific types of sensors waiting for recruitment. Next, we calculate the similarities based on the historical distribution as in the mid of Fig.2. Last, we evaluate workers' importance according to the similarities and workers' performance to recruit workers. So, our main contributions are as follows:

- We study the worker recruitment problem for multi-task data inference in Sparse Mobile Crowdsensing. From the perspective of data, we propose the data similarity and SWDTW algorithm to calculate it. Considering multiple data types, we further propose an evaluation model to obtain the similarity-based on spatial, temporal and data-type features. The similarity by such a model will be used to measure data importance.
- With the data importance, we measure the workers' utilities according to their coverage and equipped sensor types. Inspired by generalized greedy and simulated annealing, we propose a worker recruitment algorithm, WRGSA, to select suitable workers to sense important data for accurate inference. The effectiveness will be demonstrated both theoretically and experimentally.
- We conduct extensive experiments on three real-world datasets with six data types. The results verify the effectiveness of our methods on improving the data inferring accuracy under a multiple data-type scenario.

II. Related Work

Sparse Mobile Crowdsensing. Mobile Crowdsensing (MCS) [1] is an emerging method of data collection, in which people can sense various data from different areas and time slots by using their devices. After sensing data successfully, they can get corresponding rewards. Now, MCS has been profusely used in plenty of aspects of our daily life [10]–[13]. But in fact, due to the limited budget and workers, traditional MCS could not recruit enough participants which led to only a little data being sensed. To solve this problem, Sparse Mobile Crowdsensing [5] is proposed. In this paradigm, we only need to collect a little important data, then the full data is inferred according to the sensed data. The research in Sparse MCS is about recruiting workers [10], [14], improving inferring accuracy [15], [16] and so on.

Data Importance. Some research uses existing data to infer unknown data [17]–[19], but these works cannot guide us to improve the inferring accuracy when recruiting workers or assigning tasks. To solve this problem, some research focuses on evaluating the data importance [6], [20]–[22], the target of which is sensing significant areas to obtain higher inferring accuracy. To evaluate such data importance, existing works consider the data correlation in both spatial and temporal features. Wang *et al.* [6] evaluated the correlation by using the distance between different time slots then calculated the data importance based on it. Wei *et al.* [20] used the historical data and spatial distance to evaluate important sub-areas. But both of their works' disadvantage is they didn't take the scene which has multiple different data-types tasks into account.

Worker Recruitment. About worker recruitment, existing works mainly concern themselves with single-task scenes [23]–[25]. Some works [26]–[28] which are for multi-task scenarios, mostly focus on improving the workers' utilization or coverage of tasks. In contrast to the above, Wang *et al.* [6] applied the data importance and allocated tasks to workers based on it. However, they did not take the correlation between different tasks into consideration.

III. System Model

A. Concepts & Definitions

Definition 1 (Datapoint) . In this paper, we divide all MCS tasks into a set of datapoints x_i, which is denoted as X. For each datapoint $x_i = (m_{x_i}, t_{x_i}, l_{x_i})$, m_{x_i} represents the data-type of datapoint x_i, t_{x_i} represents the time slot of datapoint x_i and l_{x_i} represents the location of datapoint x_i. Furthermore, we denote the historical datapoints' set as X_{pre}.

Definition 2 (Worker). The set of workers is represented as W. For each worker $\omega = (p_\omega, c_\omega, S_\omega)$, it consists of a reliability $p_\omega \in [0, 1]$, a price c_ω, and a set S_ω which includes

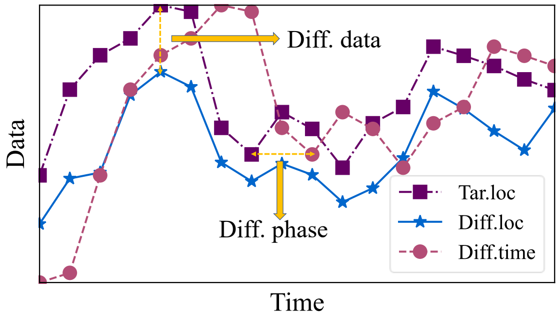


Fig. 3: Time series with different locations and time slots.

all datapoints it can sense. In this paper, we assume workers are independent of each other.

Definition 3 (Similarity) Given x_i and x_j as two different datapoints. The similarity is the correlation between x_i and x_j, which is represented as $\mathcal{G}_{i,j}$, and the range is [0,1].

For ease to explain the similarity, we sampled a location's historical data for a specified period of time (Tar.loc) with only one data type. Moreover, we also sampled the data of same period of time but different locations (Diff.loc) and the data of same location but different times (Diff.time) in Fig.3. Similarity is such an absolute criterion [9] to derive statistical inferences about the relationship between datapoints based on data and phase difference.

Definition 4 (Reliability [29]) When recruiting workers, because of the independence among them, the reliability of x_i is given by:

$$rel\left(x_i, W_{x_i}\right) = 1 - \prod_{\forall \omega_j \in W_{x_i}} \left(1 - p_{\omega_j}\right) \tag{1}$$

where W_{x_i} is a set of recruited workers. For example, when worker $\omega \in \varepsilon$, where ε is the recruited workers set, it will be added in W_{x_i} if $x_i \in S_\omega$.

Definition 5 (Interpolation Error Ratio [30]) Interpolation Error Ratio is the error ratio of inferring a missing value from a discrete set of known values by using inverse distance interpolation. For each datapoint x_i, its interpolation error ratio is expressed as $err\left(x_i\right)$. When x_i is selected, $err\left(x_i\right) = 0$.

Definition 6 (Knowledge Ratio) Knowledge Ratio is represented as the knowledge of x_i for inferring. In our paper, the knowledge ratio of the whole crowdsensing tasks equals 1 if all datapoints are sensed. For each datapoint x_i, the knowledge ratio is at most $\frac{1}{|X|}$. For ease of expression, we denote the knowledge ratio of x_i as kr_i. The formula is shown as:

$$kr_i = \frac{1}{|X|}\left(1 - err\left(x_i\right)\right) \tag{2}$$

Definition 7 (Data Quality [6]) For X which is a set of datapoints, the data quality, denoted by $Q(X)$, is defined as:

$$Q(X) = -\sum_{\forall i \in X} kr_i \log_2 kr_i \tag{3}$$

For ease of expression, when we consider the workers, $Q(X|\varepsilon)$ is the data quality based on the recruited workers ε, which is used to evaluate the importance of workers. Besides, as [6] have proven, Eq.(3) is submodular and non-decreasing.

B. Problem Formulation

In this paper, we mainly study two aspects in Sparse MCS, the one is data importance in three-dimension and the other is applying the data importance to recruit workers. For evaluating such importance, in our method, we need to calculate the similarities between datapoints. Distinctly from only considering spatial and temporal dimensions, our method will combine similarities in three dimensions which is calculated as follows:

$$\mathcal{G}_{i,j} = \alpha \cdot \mathcal{G}_{m_{x_i},m_{x_j}} + \beta \cdot \mathcal{G}_{t_{x_i},t_{x_j}} + \gamma \cdot \mathcal{G}_{l_{x_i},l_{x_j}} \tag{4}$$

where $\mathcal{G}_{m_{x_i},m_{x_j}}$, $\mathcal{G}_{t_{x_i},t_{x_j}}$, $\mathcal{G}_{l_{x_i},l_{x_j}}$ represent the similarity between the data types of x_i and x_j, the similarity between the time slots of x_i and x_j, and the similarity between the locations of x_i and x_j, respectively. Since in different situations, the weights is not equal for such three dimension similarities, we denote α, β, γ as such weights and they satisfy $\alpha+\beta+\gamma = 1$. Based on the similarity and the selected points ζ, we evaluate the importance by data quality $Q(X|\zeta)$.

In the worker recruitment stage, the scenario is more complex in that the workers may not be able to complete the task. For instance, some workers may reject the task, finish the task incorrectly, or fail to complete the task within the stipulated time constraint. Considering the above situation, [29] supposes each worker has a probability that it can successfully finish given tasks. Accordingly, we formulate the worker as $\omega = (p_\omega, c_\omega, S_\omega)$. So, we evaluate the importance of workers by the above similarity and recruit a group of them ε . Each recruited worker collects all of its datapoints with each reliability. The cost of recruited workers is given as $c(\varepsilon) = \sum_{\forall \omega_i \in \varepsilon} c_{\omega_i}$. Our goal is to maximize the data quality under the following the budget constraint which is denoted as B, and the formula is:

$$\begin{aligned} &\text{maximize} \quad Q(X|\varepsilon) \\ &\text{sub.} \quad c(\varepsilon) \leq B \end{aligned} \tag{5}$$

IV. Datapoint Importance Evaluation

A. Data Based Similarity Algorithm

Firstly, we study a simple case where we only consider a single data-type scenario at the current time, so we only need to evaluate the similarities between various locations based on historical data. As far as we know, there exist some related works on evaluating similarities between time series according to historical data. In our problem, we can change such similarities between locations as the similarity between time series. Donald et al. propose an algorithm named DTW (Dynamic Time Warping) [31]. DTW is a powerful measurement for calculating the similarity between different historical data [32]. It can evaluate the relative distance between two historical datapoints considering the phase and value difference.

Fig.4 illustrates the process of DTW in our problem. We have two locations A and B with their historical data, which are $A = a_1, a_2, a_3, ..., a_N$, $B = b_1, b_2, b_3, ..., b_N$, where N is

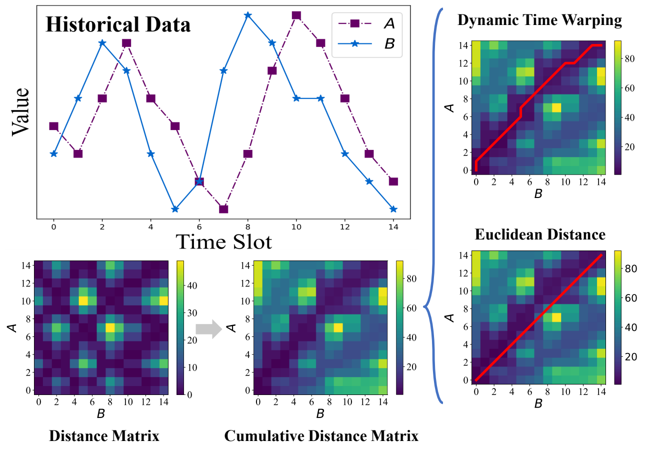


Fig. 4: An example of DTW and Euclidean distance

the length of A and B. For ease of expression, the data and corresponding datapoints have the same meaning. In Fig.4, we construct an $N \times N$ matrix which is called the distance matrix. The point (n_1, n_2) in this matrix means the distance dis_{n_1,n_2} between a_{n_1} and b_{n_2}. In DTW, such a distance is the Euclidean Distance. Next, we set Υ_{n_1,n_2} as the cumulative distance from $(1,1)$ to (n_1, n_2), DTW aims to find a path from $(1,1)$ to (N, N) which is the minimum of the cumulative distance as the distance standard between two series. Inspired by dynamic programming, Υ_{n_1,n_2} can be calculated as:

$$\Upsilon_{n_1,n_2} = dis_{n_1,n_2} + \min\{\Upsilon_{m,n-1}, \Upsilon_{m-1,n}, \Upsilon_{m-1,n-1}\} \quad (6)$$

The solution of DTW and Euclidean Distance is also shown in the right of Fig.4. As we can see, distinctly from the Euclidean Distance, DTW can calculate the similarity by taking data and phase difference into account. This is why DTW is more popular as a standard to measure time series similarity.

However, the traditional DTW has a disadvantage that it ignores the different time weights when finding such a path, which may lead to miscalculation and cause a terrible measure of similarity [33]. Moreover, in Sparse MCS, we can't promise all of the data can be sensed or precisely inferred, which means the WDTW can't be solved. So, in this subsection, we propose our method based on WDTW, which is named **Similarity-Weighted Dynamic Time Warping (SWDTW)**. The details are shown in Algorithm 1. First, considering the inferring accuracy in Sparse MCS, we set the interpolation error ratio sequences of A as $ERRA = err(a_1), err(a_2), ..., err(a_N)$ and B's as $ERRB = err(b_1), err(b_2), ..., err(b_N)$. Similarly to the traditional DTW, we also need to calculate the distance matrix. Inspired by [33], in SWDTW, dis_{n_1,n_2} is calculated as follows:

$$dis_{n_1,n_2} = \frac{2-(1-err(a_{n_1}))(1-err(b_{n_2}))}{1+exp(-g|n_1-n_2|-N/2)} \times (a_{n_1}-b_{n_2})^2 \quad (7)$$

where g is a constant. Based on the above, the Υ_{n_1,n_2} can be calculated by Eq.(6). The correlative distance between A and B, which is denoted as $\Phi_{A,B}$, is $\Upsilon_{N,N}$.

Algorithm 1 Similarity-Weighted Dynamic Time Warping

```
Input: $A, B, n$
Output: $ans$ : The relative distance between $A$ and $B$
for each $a_{n_1} \in A$ do
    for each $b_{n_2} \in B$ do
        Calculate $dis_{n_1,n_2}$ by Eq.(7)
    end for
end for
$path_{1,1} \leftarrow 0$
for each $a_{n_1} \in A$ do
    for each $b_{n_2} \in B$ do
        Calculate $\Upsilon_{i,j}$ by Eq.(6)
    end for
end for
$ans \leftarrow \Upsilon(N, N)$
```

The SWDTW can evaluate similarity considering the inferred historical data. It also prevents the same drawback of DTW, which may lead to infeasible calculations. Next, in this subsection, we can obtain the largest Φ between datapoints as the benchmark. Then we define the similarity between point i and j as follows:

$$\mathcal{G}_{i,j} = 1 - \frac{\Phi_{i,j}}{\max_{a,b \in X} \Phi_{a,b}} \quad (8)$$

B. Similarity in three dimensions

In this subsection, based on the single data-type scenario, we consider the similarity in three dimensions. To solve the first question, firstly, we set $M = \{m_{x_i} | x_i \in X\}$ as a set of all data types in X, $T = \{t_{x_i} | x_i \in X\}$ as a set of all time slots in X and $L = \{l_{x_i} | x_i \in X\}$ as a set of locations.

Then, we represent m_i as the element in M, t_i as the element in T and l_i as the element in L respectively. Because we calculate such similarities based on historical data, we make the following assumptions: we set M^*, T^*, L^* as the same set in X_{pre}, and assume that $M^* = M, L^* = L$.

For ease of explanation, we calculate the similarity between different data types as an example. To solve the problem of data-type differences (e.g., temperature, humidity, and voltage), we need to normalize the historical data for each data type as per $\tilde{d_{x_i}} = \frac{d_{x_i}-\mu}{\sigma}$, where the μ is the average and σ is the standard deviation of historical data in the same data type.

Then, we calculate the similarity between different data types. First, we need to choose corresponding data by controlling dimension. For example, when we want to calculate the $\mathcal{G}_{m_i,m_j}$, the corresponding historical datasets are represented as $D_{m_i} = \{\tilde{d_{x_i}} | x \in X_{pre}\ and\ m_x = m_i\}$ and $D_{m_j} = \{\tilde{d_{x_i}} | x \in X_{pre}\ and\ m_x = m_j\}$.

Then, we get two series from D_{m_i} and D_{m_j} in the same order. We calculate each pair's correlative distance of (m_i, m_j) as Φ_{m_i,m_j} by SWDTW, thus the similarity of data-type dimension is shown below:

$$\mathcal{G}_{m_i,m_j} = 1 - \frac{\Phi_{m_i,m_j}}{\max_{i,j \in M^*} \Phi_{i,j}} \quad (9)$$

Similarly to the above method, we calculate Φ_{l_i,l_j} from $D_{l_i} = \{\tilde{d_{x_i}} | x \in X_{pre}\ and\ l_x = l_i\}$ and $D_{l_j} = \{\tilde{d_{x_i}} | x \in$

$X_{pre}\ and\ l_x = l_j\}$, and then we calculate the similarity of spatial dimension as below:

$$\mathcal{G}_{l_i,l_j} = 1 - \frac{\Phi_{l_i,l_j}}{\max_{i,j\in L^*}\Phi_{i,j}} \tag{10}$$

In contrast to other dimensions, the temporal dimension is linear, so we can only calculate the past similarity of time slots. Because of the cyclic nature of time slots, we set the cycle length of time slots to be c, and then we set the corresponding time ${t_i}' = t_i - c$, ${t_j}' = t_j - c$. Thus we can calculate the similarity between ${t_i}'$ and ${t_j}'$ from $\Phi_{{t_i}',{t_j}'}$ using $D_{{t_i}'}$ and $D_{{t_j}'}$ as the similarity of t_i and t_j. So, the similarity of temporal dimension is formulated as:

$$\mathcal{G}_{t_i,t_j} = 1 - \frac{\Phi_{{t_i}',{t_j}'}}{\max_{i,j\in T^*}\Phi_{i,j}} \tag{11}$$

Finally, we determine the weight α, β, γ by entropy [34], [35]. We set $D_{\mathcal{G}}$ as a set including all dimensions' similarities. Then, we normalize each similarity as follows:

$$\hat{\mathcal{G}}_{i,j} = \frac{\mathcal{G}_{i,j} - \min_{\mathcal{G}\in D_{\mathcal{G}}}\mathcal{G}}{\max_{\mathcal{G}\in D_{\mathcal{G}}}\mathcal{G} - \min_{\mathcal{G}\in D_{\mathcal{G}}}\mathcal{G}} \tag{12}$$

For each dimension (time slots, locations, or data types), we calculate the similarities' probability in each dimension. For example, if we want to calculate the probabilities for pairs of data types, then we obtain the $p_{i,j,M}^{sim}$ as:

$$p_{i,j,M}^{sim} = \frac{\hat{\mathcal{G}}_{m_i,m_j}}{\sum_{i,j\in M}\hat{\mathcal{G}}_{m_i,m_j}} \tag{13}$$

Next, take data types' dimension as an example. We calculate the entropy of it, which is represented as E_M. The formula can be represented as:

$$E_M = -\frac{1}{\ln K}\sum_{k=1}^{K} p_{i,j,M}^{sim}\ln p_{i,j,M}^{sim} \tag{14}$$

where K is the pairs' size in M. Similarly to Eq.(14), we can also calculate the entropy of time slots dimension E_T and locations dimension E_L. Lastly, we calculate the value of three dimensions' weights α, β, γ as follows:

$$\begin{cases} \alpha = \frac{1-E_M}{3-(E_M+E_L+E_T)} \\ \beta = \frac{1-E_T}{3-(E_M+E_L+E_T)} \\ \gamma = \frac{1-E_L}{3-(E_M+E_L+E_T)} \end{cases} \tag{15}$$

Thus, the three-dimensional similarity is calculated by Eq.4.

C. Datapoint Importance Evaluation

In this subsection, we evaluate the data importance by using the similarity. Firstly, we modify the formula from [6] to express the interpolation error ratio as:

$$err\left(x_i\right) = \frac{\sum_{j\in S_k(x_i)}(1-\mathcal{G}_{i,j})}{k} \tag{16}$$

where $S_k\left(x_i\right)$ is a function to obtain a set of k selected datapoints with the largest similarities. As shown in Eq.(16), we select k largest similarities datapoints which belong to selected datapoints ζ. Then, we set $Q(X|\zeta)$ as the data quality based on ζ. The data quality can be calculated by Eq.(2) and Eq.(3) in that the datapoints' costs defined in this section are the same. For $x \in X - \zeta$, we can obtain the data quality variation $\triangle_x$ when x is selected as:

$$\triangle_x = Q(X|\zeta + x) - Q(X|\zeta) \tag{17}$$

Such variation is the importance of x with ζ. So, we can select important datapoints before sensing them.

V. Worker Recruitment

In this section, we consider that the workers may not be trustworthy. According to [29], we define that each worker ω has its reliability p_ω. Distinctly from only considering the datapoints, the worker recruitment mainly aims at obtaining the recruited worker set ε to maximize the data quality with the budget limitation. Before solving it, we need to prove that it is NP-hard by Theorem. 1.

Theorem 1. *The worker recruitment problem is NP-hard.*

Proof. First, we consider a simple case. A worker can only collect one datapoint in X with a single cost and 100% reliability. In this assumption, the worker recruitment problem is the same as the datapoint selection problem, which belongs to a subset selection problem, which is NP-hard. Consequently, further considering the different costs and reliability between workers, the worker recruitment problem is NP-hard. □

A. Preparation before Worker Recruitment

Based on the datapoint similarity described above, we now face a situation where each datapoint j has a reliability probability $rel(j, W_j)$ due to the unstable workers. Aiming at successfully gathering data from a datapoint with $rel(j, W_j)$, we need to construct our worker recruitment strategy after calculating the data quality that the worker acquires. In order to calculate the data quality, we first define a relative distance between datapoints with reliability as:

$$d'_{i,j} = (1 - \mathcal{G}i, j)\cdot(1 - rel(j, W_j)) \tag{18}$$

In Eq.(18), i is the origin datapoint, j is the destination. In such case, we set ${d_{i,j}}' = 1$ as the maximal distance. Such a relative distance represents the data inferring error of i from j. Then, let $S'_k(x_i)$ be a set of k datapoints with the smallest relative distance. Thus, we can represent the $err(x_i)$ as:

$$err\left(x_i\right) = \frac{\sum_{j\in S'_k(x_i)} d'_{i,j}}{k} \tag{19}$$

According to Eq.(19), inspired by [6], the knowledge ratio's formula can be modified from Eq.(2) as:

$$kr_i = \frac{1}{|X|}\left(\frac{\sum_{j\in S'_k(x_i)} rel\left(j, W_j\right)}{k} - err\left(x_i\right)\right) \tag{20}$$

Finally, we represent data quality as $Q(X|\varepsilon)$, which means the data quality under the recruited worker set ε. In this way, we can extend the datapoint importance evaluation to evaluate

Algorithm 2 Worker Recruitment by Generalized Greedy

```
Input: $B, W, X$
Output: $\varepsilon$ : The set of recruited workers
    $R^* = \{rel(x_1, W_{x_1}), rel(x_2, W_{x_2}), ..., rel(x_n, W_{x_n})\}$: The reliabilities of all data points
$\varepsilon \leftarrow \emptyset, cost \leftarrow 0$
while $cost <= B$ do
    for $\omega_i \in W - \varepsilon$ and $cost + c_{\omega_i} \leq B$ do
        compute $\frac{Q(X|\varepsilon+\omega_i)-Q(X|\varepsilon)}{c_{\omega_i}}$ as $\Delta_{\omega_i}$
    end for
    $\hat{\omega} = \arg\max\{\Delta_\omega : \omega \in W - \varepsilon\}$
    add $\hat{\omega}$ in $\varepsilon$
    $cost \leftarrow cost + c_{\hat{\omega}}$
    calculate $R$
end while
```

the contribution of each worker $\omega \in W - \varepsilon$ with existing recruited workers set ε, which is represented as:

$$\Delta_\omega = \frac{Q(X|\varepsilon + \omega) - Q(X|\varepsilon)}{c_\omega} \tag{21}$$

B. Submodular Recruitment Strategy

Next, we study the algorithm for recruiting workers based on the importance evaluation of the workers. As the proof of Theorem. 1, the problem can be seen as a subset selection problem and Eq.(3) is a submodular and non-decreasing function. The state-of-the-art strategy for such a problem is a generalized version of the greedy algorithm [36]. So, in this subsection, we recruit workers by selecting the workers with max $Q(X|\varepsilon)$. In order to maximize $Q(X|\varepsilon)$, we set Δ_{ω_i} as the heuristic value of worker ω_i. Based on this, the detail of the greedy process is shown in Algorithm 2. In each iteration, it enumerates all unselected workers and chooses the worker with highest Δ_{ω_i}. The algorithm will be teminated until all of the budget is spent.

C. Modified Strategy

Although the greedy algorithm is a powerful approach to solve this problem, it also has disadvantages because the fixed strategy may lead to a local optimum. In order to avoid it, we propose our method inspired by simulated annealing [37], which is called **Worker Recruitment by Greedy heuristic Simulated Annealing (WRGSA)**.

Algorithm 3 is the detail of WRGSA, we initiate the candidate worker recruit set ε^* by Algorithm 2. The target of this step is getting a relatively good result to speed up the cold boot of our algorithm. In each cycle, we generate a new result ε' from the candidate ε^* using Algorithm 4. Next, we compare data quality between the two solutions ε^* and ε'. If ε' is better, we will choose ε' as the new candidate recruited worker set. In contrast, if the candidate set ε^* is better, ε' will be selected as the new candidate recruited worker set with probability $\mathcal{J}_{\varepsilon^*,\varepsilon'} = exp(\frac{-(Q(X|\varepsilon^*)-Q(X|\varepsilon'))}{T})$.

Algorithm 3 WRGSA

```
Input: $B, W, X, T, T_{max}, \alpha$;
Output: $\varepsilon$ : the set of recruited workers
    $R = \{rel(x_1, W_{x_1}), rel(x_2, W_{x_2}), ..., rel(x_n, W_{x_n})\}$: the reliabilities of all datapoints
Initialize $\varepsilon^*$ by algorithm 2
while stop condition not met do
    generate a new solution $\varepsilon'$ from $\varepsilon^*$ by Algorithm 4
    if $Q(X|\varepsilon') > Q(X|\varepsilon^*)$ then
        $\varepsilon^* \leftarrow \varepsilon'$
    else
        $\varepsilon^* \leftarrow \varepsilon'$ with probability $\mathcal{J}_{\varepsilon^*,\varepsilon'}$
    end if
    if $Q(X|\varepsilon^*) > Q(X|\varepsilon)$ then
        $\varepsilon \leftarrow \varepsilon^*, T_b \leftarrow T$
        calculate $R$ according Eq.(1)
    end if
    $T \leftarrow \alpha \times T$
    if $T < 0.01$ then
        $T_b \leftarrow 2 \times T_b, T \leftarrow \min\{T_b, T_{\max}\}$
    end if
end while
```

Finally, we update the ε, which is the answer set by our algorithm, and temperature parameters. The detail is shown in Algorithm 3. In addition, for generating a new ε, Algorithm 4 includes two components: one is removing workers from ε and the other is inserting new workers into ε. The specific explanation is as follows.

Removal method: In this part, we choose probability removal to remove workers. For worker ω, we calculate its profit as $\Delta_\omega = \frac{Q(X|\varepsilon^*-\omega)-Q(X|\varepsilon^*)}{c_\omega}$. We denote $S_{min} = \{(\omega, \Delta_\omega)|\omega \in \varepsilon'\}$ as the set of each selected group of workers in ε and its profits, and then we sort all the elements in S_{min} in ascending order according to the profit Δ_ω of each element. Next, we set removal probability $\mathcal{P}_i = 2^{-i}$ of each worker ω in ε', where i is the rank of worker ω in S_{min}. Besides, to prevent duplicate selection of removed workers, let $\mathcal{L}$ be a set including removed workers. It will be initialized to be empty before removing starts, and the workers in $\mathcal{L}$ will not be selected during this execution of it. The details of it are shown in Algorithm 4.

Greedy Insertion: In order to update the workers after removing workers by probability, we recruit some workers to generate a new group of best workers. Similarly to Algorithm 2, we choose workers' set $\mathcal{W} \subseteq W - \varepsilon' - \mathcal{L}$ by greedy (the greedy heuristic value is same as Eq.(21)), and then add them in ε. The detail of it can also be seen in Algorithm 4.

Stop Condition: In this paper, we suppose it is meaningless when the algorithm can't find a better solution after a given N iterations. So, if our algorithm can't find an ε with higher data quality in the given iterations, the algorithm will be stopped.

Next, we analyze the time complexity of WRGSA. Firstly, for calculating the data quality, the time complexity of calculating data quality is $O(|X|^2 log(|X|))$. For Algorithm 4, we

Algorithm 4 Generate New Solution

```
Input: $B, W, X, cost$
    $R$: the reliabilities of all Data
    $\varepsilon^*$: a set of already recruited workers
Output: $\varepsilon'$ : the new set of recruited workers
$\varepsilon' \leftarrow \varepsilon^*, S_{min} \leftarrow \emptyset, \mathcal{L} \leftarrow \emptyset$
for each $\omega_i \in \varepsilon^*$ do
    compute $\frac{Q(X|\varepsilon'-\omega_i)-Q(X|\varepsilon')}{c_{\omega_i}}$ as $\Delta_{\omega_i}$
    $S_{min} \leftarrow S_{min} \cup (\omega_i, \Delta_{\omega_i})$
end for
sort $S_{min}$ by ascending order of $\Delta_\omega$
for each $\omega \in S_{min}$ with its rank $i$ do
    if $\omega$ is removed with $\mathcal{P}_i$ then
        $\varepsilon' \leftarrow \varepsilon' - \omega, cost \leftarrow cost + c_\omega$
        add $\omega$ in $\mathcal{L}$
    end if
end for
Select workers $\mathcal{W}$ from $\subseteq W - \varepsilon' - \mathcal{L}$ by greedy
Add $\mathcal{W}$ in $\varepsilon$
```

set $c_{min} = \min\{c_\omega|\omega \in W\}$ as the minimum cost in workers. So, its complexity is at most $O(\frac{B}{c_{min}}|W|)$. Because of the stop condition, the complexity of WRGSA is $O(N \cdot \frac{B}{c_{min}}|W|)$. Combining the data quality, the total time complexity is $O(N \cdot \frac{B}{c_{min}}|W||X|^2 log(|X|))$. Finally, we analyze the feasibility of WRGSA. As mentioned above, we utilize the greedy strategy for heuristic search. In the worst case, the effect of WRGSA is the same as generalized greedy, which can be proven as $Q(X|\varepsilon) \geq (1/2)(1-1/e) \cdot Q(X|\varepsilon_{max})$ [38], where ε_{max} is the optimal solution for recruiting workers.

VI. Experiment

In order to show the effectiveness of our proposed methods, we construct our experiments in two aspects. One aims to find out datapoints which are important for data inference; the other is recruiting a proper group of workers within the budget. All of our experiments use real datasets and we use Weighted k-Nearest Neighbor(WkNN) to infer unsensed data, by which it is easier to express the relationship between datapoints. For each unsensed datapoint x, in weighted kNN, the ground truth and inferred data is denoted as d_x and $\hat{d}_x$, $S_k(x)$ is a set of k sensed datapoints with the largest similarities. In this section, we set $k = 3$ as default.

To infer an unsensed datapoint, firstly, we need to obtain the sum of k highest similarities from sensed datapoints, which we set as Λ_x, the formula is $\Lambda_x = \sum_{y \in S_k(x)} \mathcal{G}_{x,y}$

Based on this, in this paper, all unsensed data is calculated by those sensed datapoints which are the most k-similar to it and its' corresponding weight. So, the data of the unsensed datapoint x can be inferred as $\hat{d}_x = \sum_{y \in S_k(x)} d_y \times \frac{\mathcal{G}_{x,y}}{\Lambda_x}$. Last, we choose RMSE, which is $\sqrt{\frac{1}{|X|}\sum_{i \in X}(d_i - \hat{d}_i)^2}$, to evaluate the accuracy of inference.

The last aspect of attention is how much data needs to be collected in advance. In this section, data with a cycle length of time slots, which is mentioned in Section IV, will be collected for initial similarity evaluation.

A. Dataset

The datasets we use include outdoor environmental monitoring data (**PM**, **TH**) and indoor environment data (**LVTH**)[1]. Specifically, as the outdoor datasets, PM collected PM2.5 and PM10 data from 36 subareas within an hour in Beijing. A total of 264 time slots were collected. TH obtained temperature and humidity data from 57 subareas within half an hour in the EPFL campus. The total number of time slots is 336. In the indoor dataset LVTH, four kinds of data (humidity, temperature, light and voltage data) were collected from 54 sensors in the Intel Berkeley Research Lab between February 28th and April 5th, 2004. Because of the data defection in some locations and time slots in the dataset, we selected a part of the data from the dataset and reconstructed this dataset as collecting four such data types from 33 subareas within half an hour and with a total of 384 time slots data were collected. Note that our proposed data-driven method aims to exploit the correlations between different tasks from the perspective of data similarity. So, our method is effective for multi-tasks that are intrinsically related in most cases.

B. Datapoint Selection

For datapoint selection experiments, we set each datapoint in datasets with the same cost and we assume each datapoint can be collected successfully. In conclusion, we only consider the efficiency with a given sensed ratio. In this setting, the choosing algorithm in these experiments is evaluating each datapoint's importance like in Section IV, and is sensing them by the greedy method. The greedy method is selecting the datapoints with the biggest Eq.(17). This means that this experiment is only intended to express the performance of the data quality method proposed in this paper.

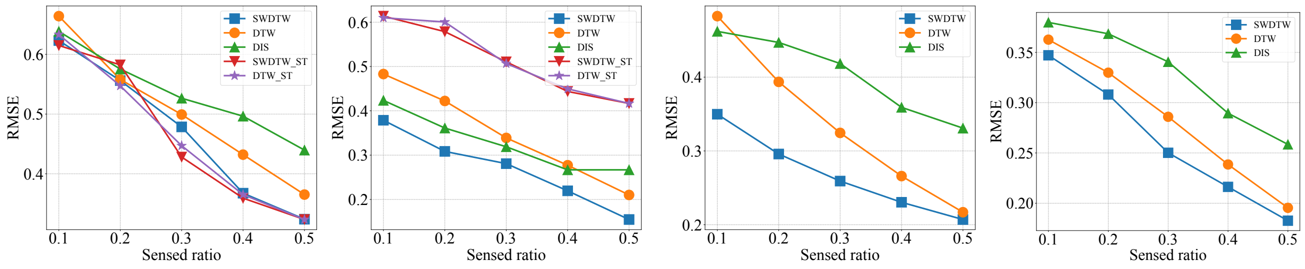
Firstly, we compare the performance of other data quality methods on our datasets. The methods' details in these experiments are as follows:

- **DIS**: This method comes from [6], and it uses the spatio-temporal distance as the standard of datapoints' similarity.
- **DTW**: This method is the DTW algorithm proposed by [31] to evaluate the similarity according historical data.
- **SWDTW**: This uses the method proposed in our paper.

Moreover, in order to show the importance of our similarity model, we also use two methods that do not consider the similarity in the data-types dimension, called **DTW_ST** and **SWDTW_ST**. In addition, in Eq.7, the constant g in dataset PM, HT and LVTH is set to be 0.25, 0.4 and 0.6 respectively.

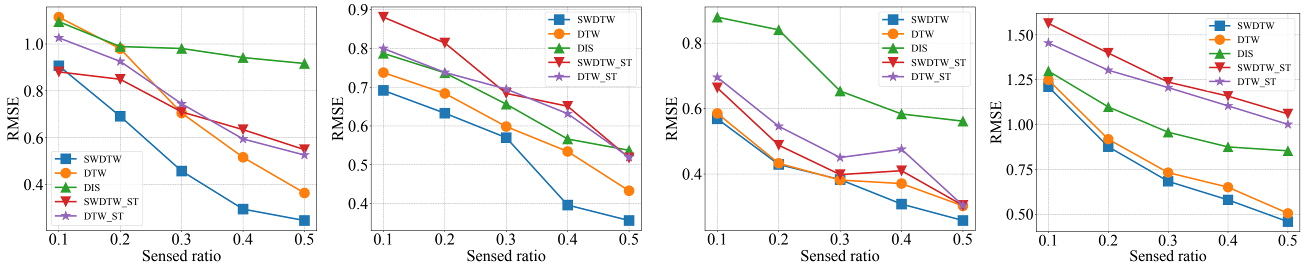
For outdoor datasets, the result of PM is shown in Fig.5(a)-(b) and the TH's is Fig.5(c)-(d). Distinctly from PM, the results of DTW_ST and SWDTW_ST are too high in HT, so we represent the results without these two methods. Lastly, the results of LVTH are contained in Fig.6.

[1] http://db.csail.mit.edu/labdata/labdata.html



(a) PM10 (b) PM2.5 (c) Humidity (d) Temperature

Fig. 5: The inferring accuracy under different sensed ratios in PM&HT.



(a) Light (b) Voltage (c) Temperature (d) Humidity

Fig. 6: The inferring accuracy under different sensed ratios in LVTH.

As we can see, as the sensed ratio increases, the value of RMSE decreases in general for each algorithm, since more sensing data can provide more similar datapoints to improve the accuracy of inference. In these experiments, the RMSE by using DTW is lower than DIS. This phenomenon told us the proposed similarity by historical data can improve the accuracy compared to just using spatio-temporal distance. On the other hand, the performance of DTW_ST and SWDTW_ST are unstable in the experiments, and both of them are the worst algorithms in most cases. This shows that the three-dimensional data similarity model in our paper is essential. The last phenomenon in this experiment is that the SWDTW performs better than DTW. It shows that the similarity algorithm in our paper is better than the traditional DTW.

C. Worker Recruitment

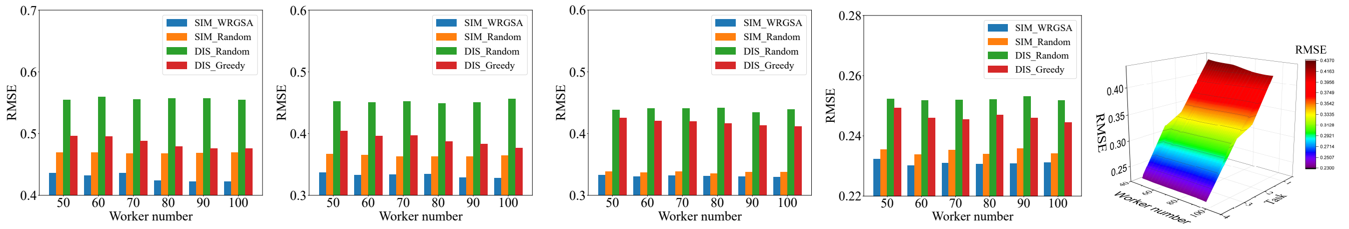
In this subsection, we verify the effect of WRGSA, which is proposed in Section V-C. Similarly to choosing a datapoint, the effect of the worker recruitment algorithm is measured by whether the algorithm can recruit a set of workers which can achieve a lower RMSE within the same budget and the same set of all workers. In addition, we evaluate the similarity in our paper with $g = 0.25$ and $N = 10$ in this subsection, and the workers' parameters are generated randomly according the datasets. In the first experiment, we evaluate our methods within a stable budget but gradually increase the workers. And we choose three methods as the baseline methods. The details of each method are shown below:

- **DIS_Random**: This method evaluates the similarity by spatio-temporal distance and selects workers randomly.
- **DIS_Greedy**: This method is proposed in [6], it evaluates the similarity by spatio-temporal distance and then utilizes the entropy to select workers greedily.
- **SIM_Random**: This method uses the similarity evaluation in Section IV but selects workers randomly.
- **SIM_WRGSA**: This method uses the similarity evaluation in Section IV and applies it by using WRGSA to recruit a group of workers.

The results of the outdoor datasets are shown in Fig.7(a)-(d), and the fitting of our algorithm's performance is shown in Fig.7(e). The indoor dataset's results are shown in Fig.8(a)-(d). Like in Fig.7(e), the fitting performance is also shown in Fig.8(e). Because of the limit of the budget, with a higher number of workers, the RMSE will decrease but tend to stabilize in a range. Moreover, as shown in the figures, the SIM_Random is better than DIS_Random. Therefore, we can conclude that in worker recruitment, the similarity proposed by our paper is also useful for inference. This phenomenon further proves the insight that our model is better than spatio-temporal distance in Sparse MCS. Furthermore, compared to other methods, the SIM_WRGSA performs much better than other algorithms. These phenomena indicate that our worker recruitment method with the similarity evaluation in this paper is more effective than other methods.

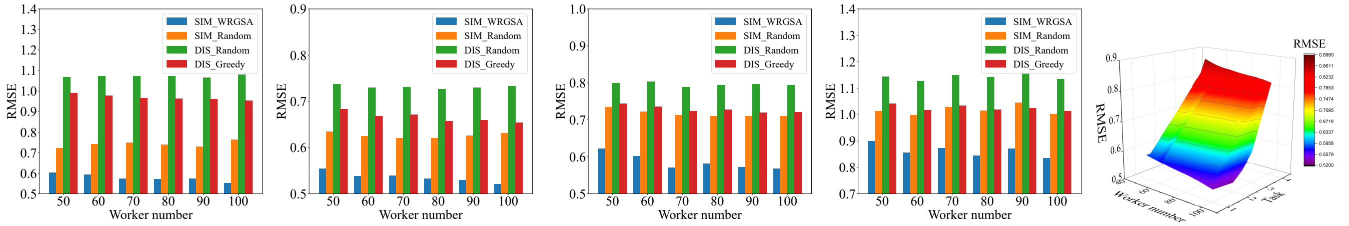
Apart from this experiment, we also compare our worker selection algorithm with other existing subset selection algorithms with stable workers but an increasing budget. In this experiment, the similarity evaluation is the same model which was proposed in our paper. We choose two published methods to verify the effectiveness of our algorithm. The algorithms' details are shown as follows:

- **WRGSA**: This is the algorithm in our paper.



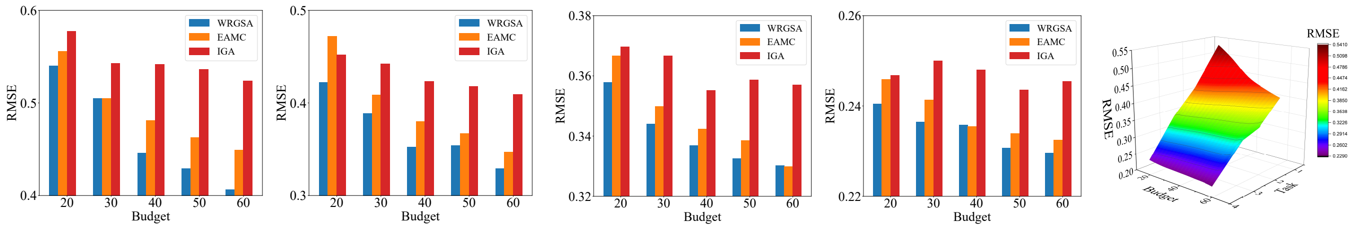
(a) PM2.5 (b) PM10 (c) Humidity (d) Temperature (e) Effect

Fig. 7: The inferring accuracy with different numbers of workers in PM&HT.



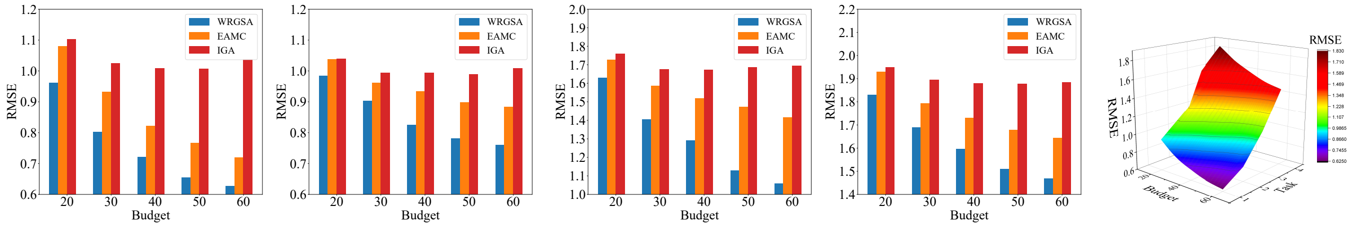
(a) Light (b) Voltage (c) Temperature (d) Humidity (e) Effect

Fig. 8: The inferring accuracy with different numbers of workers in LVTH.



(a) PM2.5 (b) PM10 (c) Humidity (d) Temperature (e) Effect

Fig. 9: The inferring accuracy under different budget constraints in PM&HT.



(a) Light (b) Voltage (c) Temperature (d) Humidity (e) Effect

Fig. 10: The inferring accuracy under different budget constraints in LVTH.

- **EAMC**: This is the modified genetic method which is proposed by [36]. This algorithm has been proven as a powerful subset selection algorithm.
- **IGA**: This is the immune genetic algorithm which is modified by [39], which has been published to solve multi-task allocation problem in MCS.

The results of this experiment are shown in Fig.9(a)-(d), and Fig.10(a)-(d) and we also show the fitting of WRGSA's performance in Fig.9(e) and Fig.10(e). As can be seen, because of the increasing budget, we can recruit more workers to sense data, which reduces the RMSE. More importantly, our method also can perform better than other algorithms. This phenomenon also proves the WRGSA's effectiveness.

VII. Conclusion

In this paper, we propose a model to calculate the similarity of different datapoints in practical Sparse MCS with three-dimensional data, such similarity is a standard for data importance evaluation. The model includes two components. One is SWDTW, which calculates the similarity between time series, which can provide a more comprehensive assessment in Sparse MCS. The other component is three dimensional similarity between data points. Finally, we apply our model to worker recruitment. Considering the reliability and the equipped sensor type of workers, we propose WRGSA for recruiting a group of workers. We conducted extensive experiments on three real-world datasets. The results of the experiments prove that our methods can select important datapoints and recruit workers to improve the inferring accuracy.

References

[1] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.

[2] D. Zhang, L. Wang, H. Xiong, and B. Guo, "4w1h in mobile crowd sensing," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 42–48, 2014.

[3] Z. Liu, S. Jiang, P. Zhou, and M. Li, "A participatory urban traffic monitoring system: The power of bus riders," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 10, pp. 2851–2864, 2017.

[4] H. Aly, A. Basalamah, and M. Youssef, "Automatic rich map semantics identification through smartphone-based crowd-sensing," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2712–2725, 2017.

[5] L. Wang, D. Zhang, Y. Wang, C. Chen, X. Han, and A. M'hamed, "Sparse mobile crowdsensing: challenges and opportunities," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 161–167, 2016.

[6] T. Wang, X. Xie, X. Cao, T. B. Pedersen, Y. Wang, and M. Xiao, "On efficient and scalable time-continuous spatial crowdsourcing," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, pp. 1212–1223.

[7] S. He and K. G. Shin, "Steering crowdsourced signal map construction via bayesian compressive sensing," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 1016–1024.

[8] L. Wang, D. Zhang, D. Yang, A. Pathak, C. Chen, X. Han, H. Xiong, and Y. Wang, "Space-ta: Cost-effective task allocation exploiting intradata and interdata correlations in sparse crowdsensing," *Acm Transactions on Intelligent Systems & Technology*, 2018.

[9] S. Lhermitte, J. Verbesselt, W. Verstraeten, and P. Coppin, "A comparison of time series similarity measures for classification and change detection of ecosystem dynamics," *Remote Sensing of Environment*, vol. 115, no. 12, pp. 3129–3152, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425711002446

[10] D. Coskun, O. D. Incel, and A. Ozgovde, "Phone position/placement detection using accelerometer: Impact on activity recognition," in *IEEE Tenth International Conference on Intelligent Sensors*, 2015.

[11] M. Gustarini, K. Wac, and A. K. Dey, "Anonymous smartphone data collection: factors influencing the users' acceptance in mobile crowd sensing," *Personal and Ubiquitous Computing*, vol. 20, no. 1, pp. 65–82, 2015.

[12] K. Lou, S. Li, F. Yang, and X. Zhang, "Advertising strategy for maximizing profit using crowdsensing trajectory data," in *International Symposium on Security and Privacy in Social Networks and Big Data*, 2020.

[13] J. Wang, Y. Wang, D. Zhang, L. Wang, C. Chen, J. W. Lee, and Y. He, "Real-time and generic queue time estimation based on mobile crowdsensing," *Frontiers of Computer Science*, 2017.

[14] T. Kandappu, A. Misra, S. F. Cheng, N. Jaiman, R. Tandriansyah, C. Chen, H. C. Lau, D. Chander, and K. Dasgupta, "Campus-scale mobile crowd-tasking: Deployment & behavioral insights," in *Acm Conference on Computer-supported Cooperative Work & Social Computing*, 2016, pp. 798–810.

[15] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, and Y. Cheng, "Truth discovery on crowd sensing of correlated entities," in *Acm Conference on Embedded Networked Sensor Systems*, 2015, pp. 169–182.

[16] X. Qiang and Z. Rong, "When data acquisition meets data analytics: A distributed active learning framework for optimal budgeted mobile crowdsensing," in *INFOCOM*, 2017.

[17] K. Xie, X. Li, X. Wang, G. Xie, J. Wen, and D. Zhang, "Active sparse mobile crowd sensing based on matrix completion," in *the 2019 International Conference*, 2019.

[18] K. Xie, J. Tian, G. Xie, G. Zhang, and D. Zhang, "Low cost sparse network monitoring based on block matrix completion," in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 2021.

[19] E. Wang, M. Zhang, X. Cheng, Y. Yang, and J. Zhang, "Deep learning-enabled sparse industrial crowdsensing and prediction," *IEEE Transactions on Industrial Informatics*, vol. PP, no. 99, pp. 1–1, 2020.

[20] X. Wei, Z. Li, Y. Liu, S. Gao, and H. Yue, "Sdlsc-ta: Subarea division learning based task allocation in sparse mobile crowdsensing," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 3, pp. 1344–1358, 2021.

[21] P. Sun, Z. Wang, L. Wu, H. Shao, H. Qi, and Z. Wang, "Trustworthy and cost-effective cell selection for sparse mobile crowdsensing systems," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 6, pp. 6108–6121, 2021.

[22] Z. Zhu, B. Chen, W. Liu, and Z. Yong, "A cost-quality beneficial cell selection approach for sparse mobile crowdsensing with diverse sensing costs," *IEEE Internet of Things Journal*, vol. PP, no. 99, 2020.

[23] W. Liu, Y. Yang, E. Wang, and J. Wu, "User recruitment for enhancing data inference accuracy in sparse mobile crowdsensing," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 1802–1814, 2020.

[24] H. Xiong, D. Zhang, L. Wang, and G. Chen, "Crowdrecruiter: Selecting participants for piggyback crowdsensing under probabilistic coverage constraint," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014.

[25] F. Li, J. Zhao, D. Yu, X. Cheng, and W. Lv, "Harnessing context for budget-limited crowdsensing with massive uncertain workers," *IEEE/ACM Transactions on Networking*, vol. 30, no. 5, pp. 2231–2245, 2022.

[26] X. Li and X. Zhang, "Multi-task allocation under time constraints in mobile crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 20, no. 4, pp. 1494–1510, 2021.

[27] J. Zhang and X. Zhang, "Multi-task allocation in mobile crowd sensing with mobility prediction," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2021.

[28] L. Wang, Z. Yu, D. Zhang, B. Guo, and C. H. Liu, "Heterogeneous multi-task assignment in mobile crowdsensing using spatiotemporal correlation," *IEEE Transactions on Mobile Computing*, vol. 18, no. 1, pp. 84–97, 2019.

[29] P. Cheng, X. Lian, Z. Chen, L. Chen, and J. Zhao, "Reliable diversity-based spatial crowdsourcing by moving workers," *Proceedings of the VLDB Endowment*, 2014.

[30] J. Gao and P. Revesz, "Voting prediction using new spatiotemporal interpolation methods," in *Proceedings of the 2006 International Conference on Digital Government Research*, ser. dg.o '06. Digital Government Society of North America, 2006, p. 293–300. [Online]. Available: https://doi.org/10.1145/1146598.1146678

[31] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, ser. AAAIWS'94. AAAI Press, 1994, p. 359–370.

[32] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: Experimental comparison of representations and distance measures," *Proc. VLDB Endow.*, vol. 1, no. 2, p. 1542–1552, aug 2008. [Online]. Available: https://doi.org/10.14778/1454159.1454226

[33] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification," *Pattern Recognition*, vol. 44, no. 9, pp. 2231–2240, 2011, computer Analysis of Images and Patterns. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132031000484X

[34] Z. hong ZOU, Y. YUN, and J. nan SUN, "Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment," *Journal of Environmental Sciences*, vol. 18, no. 5, pp. 1020–1023, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1001074206600326

[35] L. Liu, J. Zhou, X. An, Y. Zhang, and L. Yang, "Using fuzzy theory and information entropy for water quality assessment in three gorges region, china," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2517–2521, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417409007817

[36] C. Bian, C. Feng, C. Qian, and Y. Yu, "An efficient evolutionary algorithm for subset selection with general cost constraints," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, pp. 3267–3274, 2020.

[37] M. Steinbrunn, G. Moerkotte, and A. Kemper, "Heuristic and randomized optimization for the join ordering problem," *Vldb Journal*, vol. 6, no. 3, pp. 191–208, 1997.

[38] C. Qian, J.-C. Shi, Y. Yu, and K. Tang, "On subset selection with general cost constraints," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2613–2619. [Online]. Available: https://doi.org/10.24963/ijcai.2017/364

[39] X. Li and X. Zhang, "Multi-task allocation under time constraints in mobile crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 20, no. 4, pp. 1494–1510, 2021.